விரிதரவு மொழியியல்

முனைவர் ச. இராசேந்திரன்
பணிநிறை பேரசிரியர்
மொழியியல் துறை, தமிழ்பல்கலைக்கழகம்

மொழியியல் துறை
தமிழ்பல்கலைக்கழகம்
தஞ்சாவூர் 613 005

பொருளடக்கம்

இயல்/பகுதி 1: அறிமுகம்

1.1. மொழியியலும் விரிதவு மொழியியலும்

1.2. விரிதரவின் விளக்கம்

1.2.1. மாதிரியும் முன் மாதிரியும்

1.2.2. முற்று வடிவம் (Finite Size)

1.2.3. இயந்திரம் படிக்கவியலும் வடிவம் (Machine Readable Form)

1.2.4. நிலைபேறான குறிப்புரை (Standrad Reference)

1.3. விரிதரவின் முக்கியமான பண்புக்கூறுகள்

1.4. தொடக்ககால விரிதரவு மொழியியல்

1.4.1. மொழிப்பேறு (Language Acquistion)

1.4.2. எழுத்துக்கூட்டல் மரபுகள் (Spelling Conventions)

1.4.3. மொழிக்கற்பித்தல் (Language Pedagogy)

1.4.4. ஒப்பீட்டு மொழியியல் (Comparative – Linguistics)

1.4.5. தொடரியலும் பொருண்மையியலும் (Syntax and Semantics)

1.5. தொடக்க கால விரிதரவு குறித்த சாம்ஸ்கியின் கருத்து

1.6. விரிதரவு மொழியியலுக்கான வாதங்கள்

இயல்/பகுதி 2: விரிதரவு மொழியலின் இன்றைய நிலை

2.0. அறிமுகம்

2.1. எழுத்து விரிதரவு

2.1.1. பிரவுன் விரிதரவு

2.1.2. லோப் விரிதரவு

2.1.3. ஆங்கிலத்தில் ஆஸ்டிரேலியன் விரிதரவு

2.1.4. எழுதப்பட்ட நியூசிலாந்து ஆங்கிலத்தின் வெல்லிங்டன் விரிதரவு (Wellington Corpus of Written Newzealand English)

2.1.5. இந்திய ஆங்கிலத்தில் கோலாப்பூர் விரிதரவு

2.1.6. ∴ப்லோப் விரிதரவு

2.1.6. பிரிட்டிஷ் தேசிய விரிதரவு

2.1.7. அமெரிக்க தேசிய விரிதரவு

2.1.8. ஆங்கில வங்கி

2.1.9. இந்திய மொழிகளின் MIT விரிதரவு

2.2. பேச்சுத் தரவு

2.2.1. பேச்சு ஆங்கிலத்தின் லண்டன்-லண்ட் விரிதரவு

2.2.2. பேச்சு ஆங்கிலத்தின் அளவாய்வு

2.2.3. இயந்திரத்தால் படிக்கவியலும் பேச்சு ஆங்கில விரிதரவு

2.2.4. நியூசிலாந்து பேச்சு ஆய்கிலத்தின் வெலிங்டன் விரிதரவு

2.2.5. எடின்பர்க் பல்கலைக்கழக பேச்சு கால ஆணவமும் ஆங்கில மொழியின் விரிதரவும்

2.2.6. கொரியன் பேச்சு விரிதரவு

பகுதி 3: விரிதரவின் கருத்துருசார் பாகுபாடு

3.0. அறிமுகம்

3.1.பனுவலின் இனம் அடிப்படையில்

3.1.1.எழுத்து விரிதரவு

3.1.2 பேச்சு விரிதரவு

3.1.3 பேச்சு சார் விரிதரவு

3.2 தரவின் இயல்பு அடிப்படையில்

3.2.1. பொது விரிதரவு

3.2 சிறப்பு விரிதரவு

3.2.1 துணை மொழி விரிதரவு

3.2.2. மாதிரி விரிதரவு

3.2.3. இலக்கிய விரிதரவு

3.2.4. கட்டுப்படுத்தும் விரிதரவு

3.3. உரையின் வகை அடிப்படையில்

3.3.1. ஒருமொழிய விரிதரவு

3.3.2. இருமொழிய விரிதரவு

3.3.3. பன்மொழிய விரிதரவு

3.4. திட்டவரைவின் நோக்கம் அடிப்படையில்

3.4.1. குறிப்புகள்தரப்படாத விரிதரவு

3.4.2. குறிப்புகள் தரப்பட்ட விரிதரவு

3.5. பயன்பாட்டின் இயல்பு அடிப்படையில்

3.5.2. வரிசைப்படுத்தப்பட்ட விரிதரவு

3.5.3. இணையான விரிதரவு

3.5.4. நோக்கீட்டு விரிதரவு

3.5.5. ஒப்பீட்டு விரிதரவு

3.5.6. சந்தர்ப்பவாத விரிதரவு

பகுதி 4: எழுதப்பட்ட விரிதரவின் உருவாக்கம் (Generation of Written Corpus)

பகுதி 5: உரை விரிதரவு ஆய்வு (Text Corpus Processiong)

பகுதி 6: மொழித் தொழில் நுட்பத்தில் பெருந்தரவு (Corpus in Language Technology)

6.0 அறிமுகம்

6.1 மொழித்தொழில் நுட்பத்தில் பெருந்தரவின் பங்கு/மதிப்பு

6.1.1 அறிவின் மூலவளமாக பெருந்தரவு

6.1.2 மொழித் தொழில் நுட்பக் கருவிகளைத் திட்டமிடுதல்

6.1.3 மொழிபெயர்ப்புக்குதவும் ஒழுங்குமுறைகளுக்கு மூலவளமாக பெருந்தரவு

6.1.4 மனித – இயந்திர இடைமுக ஒழுங்குமுறைகளுக்கு மூலவளமாக பெருந்தரவு

6.1.5 பேச்சு தொழில்நுட்பத்தில் விரிதரவு

பகுதி 7: மொழியியலின் விரிதரவு பயன்பாடு (Corpus In Mainstream Linguistics

7.0 அறிமுகம்

7.1 அகராதியியலில் விரிதரவு **(Corpus In Lexicography)**

7.2 சொல்லியியலில் உள்ள விரிதரவு **(Corpys In Lexicology)**

7.3 இலக்கண உருவாக்கத்தில் விரிதரவின் பங்கு

7.4 பொருண்மையியல் ஆய்வில் விரிதரவு

7.5 மொழி கற்றலில் விரிதரவு (Corpus In Language Learning)

7.6 கிளைமொழி ஆய்வில் விரிதரவு (Corpus In Dialect Study)

7.7 சமுதாய மொழியியலின் விரிதரவு (Coupus In Social Linguistics)

7.8 நடையியலில் விரிதரவு (Corpus Stylistics)

7.9 உளமொழியியலில் விரிதரவு (Corpus In Psycho Linguistics)

பகுதி 8: இயந்திர மொழிபெயர்ப்பில் விரிதரவு (Corpus In Machine Translation)

8.0 அறிமுகம்

8.1நோக்கம் (The Objective)

8.2 வரலாற்றிலிருந்து கிடைக்கின்ற பாடம்

8.3 விரிதரவு அடிப்படையிலான அணுகுமுறை

8.4 விரிதரவு அடிப்படையிலான அணுகுமுறையுடன தொடர்புடைய சிக்கல்கள்

8.4 மொழிபெயர்ப்பின் விரிதரவு உருவாக்கம்

8.5. மொழிபெயர்ப்பு விரிதரவுகளில் மொழியியல் செயல்பாடுகள்

8.5.1 மொழிபெயர்ப்பு ஆய்வு

8.5.2 இருமொழிய அகராதியின் உருவாக்கம்

8.5.3 மொழிபெயர்ப்பு நிகரன்களின் பிரித்தெடுப்பு

8.5.4கலைச்சொல் தகவல் வங்கியின் உருவாக்கம்

8.5.5சொல் தேர்வு கட்டுப்பாடு

8.5.6சொல் மயக்கத்தை நீக்குதல் (Sissolving Lexical Ambkguity)

இலக்கண பொருத்தம் (Grammartical Mapping)

இறுதியுரை

பகுதி 9:முடிவுரை

அறிமுகம்

1.2. மொழியியலும் விரிதவு மொழியியலும்

மொழியியல் என்பது மனித நாகரீகத்தின் மிக பழமையான பாடமாகும். நாம் பல்லாயிரக்கணக்கான ஆண்டுகளாக மனிதமொழியின் வேறுபட்ட நோக்குகளை ஆய்ந்து வருகின்றோம். பல நேரங்களில் மனித மொழி அறிவு, புலனறிவு மற்றும் கருத்துப் பரிமாற்றம் இவை குறித்து ஆய்வதற்கு முக்கியத்துவம் தரப்பட்டது. பல நூற்றாண்டுகளாக இப்பாடம் மனித அறிவின் பிற கிளைகளுடன் அதன் கருத்துருசார் தொடர்பை நிறுவிப் படிப்படியாக அதன் முழு வடிவில் வளர்ந்துள்ளது. இப்போது, இன்றைய நூற்றாண்டின் தொடக்கத்தில் இது எவ்வாறு மொழியின் வேறுபட்ட நோக்குகளைப் பற்றிய கோட்பாடுகள் சீரான மனித மொழி வெளிப்பாடுகளில் உருப்படுத்தம் செய்யப்பட்டுள்ள மொழியின் உண்மையான பயன்பாட்டிற்குச் சான்றாக மெய்படுத்தம் செய்யப்பட்டுள்ளது என்பதை ஆய்வதில் தன் கவனத்தைத் திருப்பியுள்ளது.

மொழி ஆய்வின் இப்புதிய திசை மரபு மொழியியலுக்குக் கூடுதல் பரிமாணத்தைத் தந்துள்ளது. மொழியியல் ஆய்வுக்குப் புதிய கருவிகளின் குழுமத்தையும் மொழியின் உண்மையான பயன்பாட்டின் எடுத்துக்காட்டுகளைச் சேகரிக்கும் தொழில் நுட்பங்களையும் நல்கியக் கணிப்பொறித் தொழில் நுட்பத்தின் அறிமுகத்தால் இது சாத்தியமானது. இப்புதிய அணுகுமுறையின் அறிமுகம் மொழியிலுக்கும் மனிதவினத்திற்கும் இரண்டு வழிகளில் நன்மைபுரிந்துள்ளது. முதலாவது இது மொழி மற்றும் மொழிப்பயன்பாடு பற்றிய பழைய கோட்பாடுகள் தொடர்ந்து பயன்படுத்த உகந்ததா என்பதைத் தெரிந்துகொள்ள உதவுகின்றது. இரண்டாவது இது சீரான மனிதச் செயல்பாடுகளில் மொழிச் சான்றின் மற்றும் தகவலின் நேரடியான பயன்பாட்டிற்கு அனுகூலங்கள் தருகின்றது. இது சரியான திசை, மாறுபாடு மற்றும் பயன்பாடு என்பனவற்றின் குறைவால் சிலகாலம் பாதிக்பட்டுள்ள மொழியலின் மறுமலர்ச்சிக்கும் நிலைநிற்புக்கும் உதவும்.

சில காலகமாகக் கணிப்பொறித் தொழில்நுட்பத்தின் கண்டுபிடிப்பு மற்றும் முன்னேற்றம் மொழியிலுக்கு ஒரு புதிய பரிமாணத்தைத் தந்துள்ளது; மொழியைப் புலனறிவுடன் தொடர்புடைய மனிதக் கருத்துப்பரிமாற்றத்தின் கருவியாகப் பார்ப்பதை நோக்கமாகக் கொண்ட செயற்கை அறிவுநுட்பத்தின் (artificial intelligence) பகுதியாகக் கணிப்பொறி மொழியியல் (computational linguistics) என்ற பதிய பாடம் உருவாகியுள்ளது. விரிதரவு மொழியியல் கணினிமொழியியலின் முக்கியமான கிளையாகப் புள்ளியியலில் பயன்படுத்தப்படும் நெறிகள் மற்றும் உத்திகளைப் பின்பற்றி மிக ஒழுங்கான வழியில்

சேகரிக்கப்பட்டுள்ள மொழிப் பயன்பாட்டின் அனுபவவாதச் சான்றின் பெரும் அளவைத் தரும் செயல்பாட்டில் முக்கியப் பங்களிப்பு செய்கின்றது. மேலும் இது விரிதரவை ஆயவும் மனித மொழியைப் புரிந்துகொள்ளவும் மனித அறிவியலின் பல களங்களில் பயன்படுத்தவும் கணினி மொழியியலிலும் செயற்கை அறிவுநுட்பத்திலும் தேவையானதாகக் கருதப்படும் தகவல்களைப் பெற மிக நுட்பமான உபாயங்களின் குழுமத்தைத் தருகின்றது.

எவ்வாறு மக்கள் மொழியை ஒரு கருவியாகப் பயன்படுதி தங்களுக்குள் கருத்துப்பரிமாற்றம் செய்கின்றனர் என்பதைப் புரிந்துகொள்ள வலுவான புலனறிவு மற்றும் மொழியியல் சார்ந்த நோக்கம் இருக்கின்றது. மனிதவினத்துடன் திறமையாக மொழி ஊடாட்டங்கள் செய்யவியலும் அறிவுநுட்பமுள்ள கணிப்பொறி ஒழுங்குமுறைகளை உருவாக்க தொழிநுட்டபம் அடிப்படையிலான நோக்கம் இருக்கின்றது. இந்த நோக்கங்களுடன் கணிப்பொறி அறிவியலார்களும் மொழியலார்களும் ஒன்றாக இணைந்து இயந்திர மொழிபெயர்ப்புக்கும் தகவல் பிரித்தெடுப்புக்கும் மொழி புரிதலுக்கும் மொழி உருவாக்கத்திற்கும் ஒழுங்குமுறைகளை உருவாக்கியுள்ளனர். இவ்வேலைகளுக்கு அவை சீரான மற்றும் அரிதான பண்புக்கூறுகளைக் கொண்ட இயற்கை மொழியை அனுபவவாத அடிப்படையில் புரிந்துகொள்ள வேண்டியது அவசியமாகும். இங்கு விரிதரவு தவிர்க்க முடியாததாக மாறுகின்றது; ஏனென்றால் விரிதரவு அனுபவாத தரவுமையங்களின் சேகரிப்புக்குள் அறியவியலும் மொழியின் பெரும்பாலான பண்புக்கூறுகளை வெளிக்கொணர திறன்களைக் கொண்டுள்ளது.

தற்போது பல எண்ணிக்கையிலான மக்கள் கணிப்பொறியில் பல்வேறு வகைப்பட்ட மொழியியல் தகல்களை நடைமுறைப்படுத்துவதில் ஈடுபட்டுள்ளனர்; ஏனென்றால் கணினி மொழியியலின் மற்றும் இயற்கை மொழியாய்வின் (natural language processing) நோக்கங்கள் மொழியின் பண்புக்கூறுகளை முடிந்த அளவு கணினி அடிப்படையில் பண்பாக்கம் செய்யவேண்டும் என்பதாகும். மேலும் இயற்கை மொழியியல் ஆய்வுக் களத்தில் செய்யப்பட்டுள்ள ஆராய்ச்சிகள் ஒரு இயற்கை மொழியின் மொழியியல் பகுப்பாய்வில் வெளிப்படுத்தப்படும் உள்ளறிவிலிருந்து நன்மையடைந்துள்ளன. எடுத்துக்காட்டாக, ஒரு கணிப்பொறியால் தானாக ஒரு எளிய வாக்கியத்தைப் பொருள்கோள் செய்ய ஒழுங்குமுறையை உருவாக்குபவர் இம்மாதிரி வாக்கியங்களின் முந்தைய மொழியியல் பகுப்பாய்வு பயனுள்ளதாகவும் அறிவுபுகட்டுவதாகவும் இருப்பதைக் காண்பார். எனவே ஒரு விரிதரவுக்குள் சேகரிக்கப்பட்டுள்ள மொழி அலகுகளின் வர்ணனைகளும் தன்மைகளும் மொழித் தொழில் நுட்பத்திற்கு (languge technology) மிக முக்கியமானவைகளாகும். விரிதரவிலிருந்து எடுக்கப்படும் புதிய தகவல்கள் கணினி மொழியியலுக்குப்

பயன்படுவதுடன் முக்கிய மொழியியல் களத்திற்குப் பயன்படும் மொழியின் வருணனை மற்றும் புரிந்துகொள்ளல் பற்றிய மதிக்கத்தக்க உள்ளறிவுகளையும் தருகின்றது.

மொழியில் மொழி ஊடாட்டம், கருத்துப்பரிமாற்றம் மற்றும் புரிந்துகொள்ளல் என்ற பல்வேறு களங்களில் வெளிப்படுத்தப்பட்டுள்ள ஒரு இயற்கை மொழியின் பயன்பாட்டின் எல்லா திசைகளையும் உள்ளடக்கும் விரிந்த பரப்பைக் கொண்ட பல்பரிமணக் களமாகும். மொழி ஆய்விலும் பயன்பாட்டிலும் விரிதரவின் அறிமுகம் மைய மொழியிலுக்குப் புதிய பரிமாணங்களை உட்படுத்தியுள்ளது. கொள்கை அடிப்படையில் விரிதரவு மொழியியல் (corpus linguistics) இயந்திரத்தால் படிக்கவியலும் உரைகளின் மிகப்பெரிய சேகரிப்புகளின் வழி மொழியியல் நடப்புகளை ஆய்வதை நோக்கமாகக் கொண்ட அணுகுமுறையாகும். இவ்வணுகுமுறை பல ஆய்வுக்களங்களில் பயன்படுத்தப்படுகின்றது: ஒரு மொழியின் வருணனை ஆய்விலிருந்து மொழித் தொழில் நுட்பம் மற்றம் கல்வி வரையிலான எல்லா ஆய்வுகளிலும் பயன்படுத்தப்படுகின்றது.

விரிதரவு மொழியியலின் உள்ளுறையும் கருத்துரு விரிவாகப் பலவற்றை உள்ளடக்கியதாகும். இது நம்பகமான பேச்சு மற்றும்/அல்லது எழுத்து உரை மாதிரிகளின் எந்தக் கணிசமான பகுதியின் முழுமையான ஆய்வையும் குறிப்பிடும். இது மொழியின் ஒருகால மற்றும் இருகால நோக்குகளை வருவரைவு செய்யும் இலக்கிய உரைகள் மற்றம் செய்தித்தாள் கட்டுரைகள் இவற்றின் சேகரிப்புகளை உள்ளடக்கிய தற்கால மொழியின் பெரிய அளவிலான இயந்திரத்தால் படிக்கவியலும் விரிதரவுகளை உட்படுத்தும்.

மின் விரிதரவு (electronic corpus) என்பது புதிய ஒன்றாகும். இக்களம் அரை நூண்ற்ாண்டு வரலாறு உடையது. எது ஒரு விரிதரவாகக் கருதப்பட வேண்டும், அது எவ்வாறு திட்டமிடப்படவேண்டும், எவ்வாறு உருவாக்கப்படவேண்டும், வகைப்படுத்தப்படவேண்டும், ஆயப்படவேண்டும் மற்றும் பயன்படுத்தப்படவேண்டும் என்பதில் பொதுவான கருத்து இல்லை.

1.2. விரிதரவின் விளக்கம்

அனுபவவாத ஆய்வுகள் எந்த எழுதப்பட்ட அல்லது பேசப்பட்ட உரைகளையும் பயன்படுத்தி அதன் மேல் செய்யப்படுவதாகும். இம்மாதிரியான தனிப்பட்ட உரைகள் பல வகையான இலக்கிய மற்றும் மொழியியல் ஆய்வுகளுக்கு அடிப்படையாக அமைகின்றது. எடுத்துக்காட்டாக ஒரு செய்யுளின்

அல்லது ஒரு புதினத்தின் நடையியல் ஆய்வு அல்லது ஒரு தொலைக்காட்சிப் பேச்சின் உரையாடல் ஆய்வு என்பதைக் கூறலாம். ஆனால் அனுபவாத மொழியியலின் ஒரு வடிவுக்கு அடிப்படையாக அமையும் ஒரு விரிதரவு என்பதன் கருத்துச்சாயல் உரைகளின் குறிப்பிட்ட பகுதியை பரிசோதிக்கும் வழிகளிலிருந்து பல அடிப்படையான வழிகளில் வேறுபடும். கொள்கை அடிப்படையில் ஒரு உரைக்குக் கூடுதலான எந்தச் சேகரிப்புத் தொகுதியும் விரிதரவு எனப்படும்: விரிதரவு என்று தமிழில் அழைக்கப்படும் corpus என்பது body (தேகம்/உடல்) என்பதன் இலத்தின் சொல்லாகும்; எனவே விரிதரவு என்பதை எந்த ஒரு உரையின் உடல் என்று கூறலாம். ஆனால் விரிதரவு என்ற சொல் தற்கால மொழியியல் என்ற சூழலில் பயன்படுத்தப்படும் போது, எளிய விளக்க உரையைத் தாண்டிக் கூடுதல் சிறப்பு அர்த்தங்களைக் கொண்டிருக்கும் போக்கு காணப்படுகின்றது. இதைப் பின்வரும் நான்கு முக்கியத் தலைப்புகளின் அடிப்படையில் கருதலாம்:

1. மாதிரி மற்றும் முன்மாதிரிகள் (Sampling and representativeness)
2. முற்று வடிவம் (Finite Size)
3. இயந்திரம் படிக்க இயலும் வடிவம் (Machine readable form)
4. நிலைபேறான குறிப்புரை (Standard reference)

1.2.3. மாதிரியும் முன் மாதிரியும்

மொழியியலில் நாம் ஒரு குறிப்பிட்ட உரை அல்லது நூலாளர் என்பதை விட ஒரு மொழியின் முழு வேறுபாட்டில் ஆர்வம் காட்டுவோம். இந்நேர்வுகளில் நமக்குத் தரவு சேகரிப்பில் இரண்டு விருப்புகள் உள்ளன: முதலாவது, நாம் அந்த வகையின் ஒவ்வொரு கூற்றையும் ஆய இயல வேண்டும்; அல்லது இரண்டாவது, நாம் அந்த வகையின் ஒரு சிறிய மாதிரியை உருவாக்க இயல வேண்டும். முதலாவது விருப்பு, செயல்பாட்டு அடிப்படையில் இயலாததாகும். உரைகளின் தொகை மிகப்பெரியதாகும். ஆங்கிலம், தமிழ் போன்ற மொழிகளில் இது அதிகரித்துக் கொண்டே செல்கின்றது. இது எல்லை இல்லாதது என்று கூறலாம். எனவே நாம் ஒரு மொழி வகையின் விரிதரவை உருவாக்க ஒரு மாதிரியை மட்டும் கருத்தில் கொள்வோம். எடுத்துக்காட்டாக, நாம் பலவகைப்பட்ட நாவல்கள் மற்றும் எழுத்தாளர்களின் நாவல்கள் இவற்றின் விரிந்த எல்லைக்கு உட்பட்ட ஒரு முன் மாதிரியைத் தேர்ந்தெடுப்போம்.

1.2.4. **முற்று வடிவம்** (Finite Size)

மாதிரியைப் போன்று விரிதரவு என்பது ஒரு முற்றுவடிவத்தில் பனுவலின் உடலைக் குறிப்பிடும் போக்கு இருக்கின்றது. எடுத்துக்காட்டாக, ஒரு விரிதரவின் வடிவம் ஒரு மில்லியன் சொற்களில் உருவாக்கப்படலாம். இது உலக பொதுமையானது அல்ல. எடுத்துக்காட்டாக, பிர்மிங்காம் பல்கலைக்கழகத்தில் ஜாண் சின்கிளேயரின் COBUILD குழு 'monitor corpus' என்ற உரைகளின் சேகரிப்பின் உருவாக்கத்திலும் ஆய்விலும் ஈடுப்பட்டுள்ளனர். இது திறந்த முடிவற்ற விரிதரவாகும். இம்மாதிரியான விரிதரவு அகராதி உருவாக்கத்திற்கு முக்கியமானதாகும். லாங்கமானின் COBUILD ஆங்கில அகராதி இந்த விரிதரவால் உருவாக்கப்பட்டதாகும்.

1.2.3. **இயந்திரம் படிக்கவியலும் வடிவம்** (Machine Readable Form)

இன்றைய காலகட்டத்தில் விரிதரவு என்றாலே அது இயந்திரம் படிக்கக் கூடிய வடிவில் இருக்க வேண்டும் என்பதை உட்கோள் செய்யும். பல ஆண்டுகளாக விரிதரவு என்பது அச்சிடப்பட்ட உரைகளைக் குறிப்பிடுவதாக இருந்தது. இப்பொழுது இந்நிலை மாறிவிட்டது. சில பெருந்தரவுகள் அச்சு வடிவத்திலும் இருக்கிறது. எடுத்துக்காட்டாக A Corpus of English Conversation (Svartvik and Quirk, 1980) என்பது அச்சடிக்கப்பட்ட வடிவில் இருக்கின்றது. இந்த விரிதரவு மூல வடிவான London–Lund–Corpus என்பதை உருப்படுத்தம் செய்கின்றது. London-Lund Corpus, இயந்திரம் படிக்க இயலும் வடிவத்துடன் இருப்பதுடன் புத்தக வடிவிலும் இருக்கிறது. Lancaster/IBM Spoken English Corpus பேச்சு விரிதரவுக்கு எடுத்துக்காட்டாகும். இயந்திரம் படிக்க இயலும் வடிவம் பல ஆய்வுகளுக்கு பயன் உள்ளதாக அமைகின்றன. இது அச்சு வடிவில் அல்லது பேச்சு வடிவில் உள்ள விரிதரவைக் காட்டிலும் மிக பயனுள்ளதாக இருக்கின்றது.

1.3.4. **நிலைபேறான குறிப்புரை** (Standrad Reference)

இது விரிதரவின் வரையறை விளக்கத்திற்கு முக்கியமான பகுதியாக இல்லாவிட்டாலும் ஒரு விரிதரவு அது உருப்படுத்தம் செய்யும் மொழி வகையின் நிலைபேறு குறிப்புரையாக இருக்கும் என்பது எதிர்பார்ப்பாகும். எடுத்துக்காட்டாக, அமெரிக்க ஆங்கில மொழியின் Brown Corpus, எழுத்து பிரிட்டன் ஆங்கிலத்தின் LOB Corpus, பேச்சு பிரிட்டன் ஆங்கிலத்தின் London–Lund Corpus என்பன ஆய்வாளர்களால் பரவலாகப் பயன்படுத்தப்படுகின்றன.

இவ்வாறு இக்கால மொழியியலில் ஒரு விரிதரவு என்பது ஒரு குறிப்பிட்ட பனுவலாக இல்லாமல் கருதப்பட்ட மொழி வகையின் பிரதிநிதியாக இருக்கத்தக்க விதத்தில் மாதிரியாக்கம் செய்யப்பட்டு இயந்திரத்தால் படிக்க இயலும் முற்றுப்பெற்ற வடிவம் என்று வருணிக்கத்தக்க விதத்தில் இருக்கின்றது.

1.4. விரிதரவின் முக்கியமான பண்புக்கூறுகள்

கோட்பாடு அடிப்படையில் corpus என்பததை பின்வருமாறு விளக்கலாம்: **C**apable **O**f **R**epresenting **P**otentially **U**nlimited **S**election of **T**exts. அதாவது விரிதரவு ஆற்றல் அடிப்படையில் எல்லையற்ற உரைகளின் தேர்வை உருப்படுத்தம் செய்ய இயலும். It is **C**ompatible to computers, **O**perational in research application, **R**epresentative of source language, **P**orcessable by men and machine, **U**nlimitted in data, and **S**ystematic in formation and representation. அதாவது விரிதரவு கணிப்பொறியுடன் இயைபுள்ளது, ஆய்வுப் பயன்பாட்டில் செயன்மைபுரிவது, மூலமொழியின் பிரதிநிதி, மனிதனாலும் பொறியாலும் பகுத்தாயத்தக்கது, தரவில் எல்லையில்லாதது மற்றும் ஆக்கத்திலும் உருப்படுத்தத்திலும் ஒழுங்குமுறையானது.

விரிதரவின் முக்கிய பண்புக்கூறுகளாகப் பின்வருவனவற்றைக் குறிப்பிடலாம்:

அளவு: ஒரு விரிதரவு பேச்சுவடிவிலோ எழுத்து வடிவிலோ இருக்கும் மிகப் பெரிய அளவு தரவைக் கொண்ட மிகப்பெரிய உருவுள்ளது. உரு அதன் உடலை உருவாக்கும் உறுப்புக்களின் கூட்டுத்தொகையாகும்.

பண்பு 'நம்பகத்தன்மை' : எல்லா விஷயங்களும் பேச்சின் மற்றும் எழுத்தின் மெய்யான இயல்பான பயன்பட்டிலிருந்து பெறப்படவேண்டும். இங்கு மொழியியலாரின் பங்களிப்பு தேவை. அவர் தரவு பரிசோதனைக் கட்டுப்பாடுகளோ செயற்கையான சூழல்களோ உள்ள கருத்துப்பரிமாற்றத்திலிருந்து பெறப்படாமல் சாதாரண கருத்துப்பரிமாற்றத்திலிருந்து பெறப்பட்டதா என்று பரிசோதிக்க வேண்டும்.

பிரதிநிதித்துவம்: இது விஷயங்களின் விரிந்த பரப்பெல்லையிலிருந்து பெற்ற மாதிரிகளை உட்படுத்தும். இது ஒரு மொழியின் கூடுதலான மொழிப் பண்புக்கூறுகளை பிரதிநிதித்துவம் செய்யவேண்டி இது எல்லாத் துறைகளிலிருந்தும் களங்களிலிருந்தும் சரிநிகரான விஷங்களைக் கொண்டிருக்கின்றது. இதின் மீது திட்டமிடப்படும் எதிர்கால ஆய்வு ஒரு மொழியை பிரதிநிதித்துவம் செய்யும் விரிதரவிலிருந்து பெறப்பட்ட தகவலின் பரிசோதனையையும் வேண்டும்.

எளிமை: விரிதரவு எளிய வெளிப்படையான உரைகளைக் கொண்டிருக்கும். அதாவது நாம் பனுவலுக்குள் எந்தக் கூடுதலான மொழித் தகவலும் அடையாளப்படுத்தப்படாத எழுத்துக்களின் தொடர்ச்சியான கோர்வையை எதிர்பார்கின்றோம். ஒரு எளிய வெளிப்படையான உரை மொழிசார்ந்த மற்றும் மொழிசாராத தகவல்களின் பல வகைகளைக் கொண்டிருக்கும் எந்த அடையாளப்படுத்தலுக்கும் எதிரானதாகும்.

சமத்துவம்: விரிதரவில் பயன்படுத்தப்படும் மாதிரிகள் நிகரான அளவில் இருக்கவேண்டும். இருப்பினும் இது விவாதத்திற்குரிய சிக்கல்; எவ்விடத்திலும் ஏற்றுக்கொள்ளப்படமாட்டாது. மாதிரிசார் மாதிரி (sampling model) விரிதரவைக் கூடுதல் பிரதிநிதித்துவம் செய்யவும் பல்பரிமாணம் உடையதாகச் செய்வும் வேண்டி போதுமான அளவு மாறும்.

மீளப்பெறும் தன்மை: தகல்கள், எடுத்துக்காட்டுகள் மற்றும் நோக்கீடுகள் என்பன பயன்படுத்துவோர்களால் விரிதரவிலிருந்து எளிதாக மீளப்பெறத்தக்கதாக இருக்கவேண்டும். இது பயன்படுத்துவோர்களுக்கு வேண்டி கணினியில் மின்வடிவில் தரவைப் பாதுகாக்கும் தொழிநுட்பத்தில் கவனம் செலுத்தும். இன்றைய தொழில் நுட்பம் தனிநபர் கணிப்பொறியில் விரிதரவை உருவாக்கவும் நாம் தேவைப்படும் போது எளிதாகத் தரவை மீளப்பெறவும் சாத்தியமானதாகச் செய்துள்ளது.

பரிசோதனை செய்யும் தன்மை: விரிதரவு எந்த வகையிலான அனுபவாத ஆய்வுக்கும் பரிசோதனைக்கும் உட்படுத்தப்படவேண்டும். நாம் தரவுமையத்தை உள்ளுணர்வு மதிப்பீட்டால் செய்யப்படும் உற்றுநோக்கல்களை பரிசோதிக்கவோ மறுக்கவோ வேண்டி எந்த புலன்விராரணைக்கும் பயன்படுத்தலாம்.

விரிவுபடுத்தும்தன்மை: விரிவுபடுத்தல் சீர்மையை அதிகரிக்கவேண்டும். இது விரிதரவைக் காலப்போக்கில் மொழிக்குள் காணப்படும் மொழி மாற்றங்களை பதிவுசெய்வற்கு இணையானதாகச் செய்யும். காலப்போக்கில் புதிய மொழித் தரவை முறையாக அதிகரிப்பதால் இருகால மொழி ஆய்வுகளுக்கு வேண்டி தகவல்களைச் சேகரிப்பதற்கும் வாழ்க்கையிலும் சமுதாயத்திலும் உள்ள மாற்றங்களைத் தடை செய்யவேண்டிய மொழி உளவுக்கூறுகளை வெளிப்படுத்திக் காட்டவும் வேண்டி விரிதரவு வரலாற்றுப் பரிமாணத்தைப் பெறும்.

1.4. தொடக்ககால விரிதரவு மொழியியல்

'தொடக்க கால விரிதரவு மொழயியல்' (Early Corpus Linguistics) என்ற கலைச் சொல்லை நாம் சாம்ஸ்கியின் வரவுக்கு முன்னால் உள்ள மொழியியல் என்பதை விளக்கப் பயன்படுத்தலாம். களமொழியியல் (Field Linguistics) மற்றும் அமைப்புமொழியியலார் மரபைச் சார்ந்த பிந்தைய மொழியியலார்கள் பயன்படுத்திய அடிப்படை நெறிமுறையை எந்த ஐயமும் இன்றி விரிதரவு அடிப்படையிலானது என்று கூறலாம். ஆனால் இக்கால கட்டத்திலிருந்தே 'விரிதரவு மொழியியல்' என்ற கலைச்சொலைப் உரைகளிலும் ஆய்வுகளிலும் காணலாம் என்று அர்த்தம் அல்ல. 'ஆரம்ப கால விரிதரவு மொழியியல்' என்ற கலைச்சொல் இப்படைப்புகளை வகைப்பாடு செய்யவேண்டி உருவாக்கப்பட்டதாகும். இது சில வழிகளில் சாம்ஸ்கிக்கு முந்தைய மொழியியல் நெறிமுறையைக் குறிப்பதுடன் அது உறவு வைத்துள்ள புதிய விரிதரவு மொழியியல் நெறிமுறையுடன் தொடர்புகொண்டுள்ளது. இச்செய்தியை எடுத்துக்காட்டவும் 1950-களுக்கு முன்னர் மொழியியலில் விரிதரவின் பரந்த பயன்பாட்டை நிறுவவும் கீழே அக்காலகட்டத்திய விரிதவு அடிப்படையிலான ஆய்வுகளின் சுருக்கம் தரப்பட்டுள்ளன:

1.4.1. மொழிப்பேறு (Language Acquistion)

மொழியியல் 19-ம் நூற்றாண்டிலும் 20-ம் நூற்றாண்டின் தொடக்க காலத்திலும் விரிதரவு அடிப்படையிலான வர்ணனையை கையாண்டது. ஆரம்பகாலத்தில் (ஏறக்குறைய1876-இலிருந்து 1926-வரை) மொழிப்பேறு குறித்த ஆய்வுகள் குழந்தைகளின் பேச்சைப் பதிவுசெய்யும் செய்யும் குறிப்பேடு அணுகுமுறையாக இருந்தன. இவை குழந்தைகளின் பேச்சைப் பதிவுசெய்யும் பெற்றோர்களின் கவனமாகத் தொகுக்கப்பட்ட குறிப்பேடுகள் அடிப்படையில் அமைந்தன. பிரயர் (Preyer, 1988) மற்றும் ஸடர்ன் (Stern, 1924) என்ற ஆய்வாளர்களுக்கு அடிப்படையக அமைந்த பழைய விரிதரவுகள் இன்றைய மொழிப்பேறு ஆய்வுகளில் தரவு மூலமாகப் பயன்படுத்தப்படுகின்றது. விரிதரவு சேகரிப்பு தொடர்ந்து நடைபெற்றது மற்றும் குறிப்பேடு பதிவு செய்யும் காலத்தைத் தாண்டி பல்வேறுவிதமாக மாறியது. குழந்தைகளின் மொழியை உருப்படுத்தம் செய்யும் கூற்றுக்களின் விரிதரவைச் சேகரிக்கப் பயன்படுத்தப்படும் மாதிரி உத்தியின் அடிப்படையில் மொழிப்பேறு ஆய்வுகளை நீளப்பாங்கு ஆய்வுகள் (longitudinal studies) மற்றும் பெரும் மாதிரி ஆய்வுகள் (large sample studies) என்று பிரிக்கலாம்.

பெரும் மாதிரி ஆய்வுகள் 1927-இலிருந்தது 1957-வரை உள்ள காலகட்டத்தை உள்ளடக்கும். மொழிப்பேறு வளர்ச்சியின் நியமங்களை நிறுவும் நோக்கத்துடன் ஆய்வுகளுக்கான விரிதரவுகள் பெரும்

எண்ணிக்கையிலான குழந்தைகளிடமிருந்து பெறப்பட்டன (எடுத்துக்காட்டாக மெகார்த்தி (McCarthy 1954)).

நீளப்பாங்கு ஆய்வுகள் 1957-இலிருந்து இன்றுவரை ஆதிக்கம் செலுத்துகின்றன. இவைகளும் கூற்றுக்களின் சேகரிப்பு அடிப்படையில் அமைந்தவை; ஆனால் காலம் தழுவிய தரவுமூலமாக மூன்று குழந்தைகளைச் சுற்றிய சேகரிப்பு பயன்படுத்தப்பட்டது. பிரவுண் (Brown, 1973) மற்றும் புளூம் (Bloom, 1970) என்பவர்களின் ஆய்வுகள் நீளப்பாங்கு ஆய்வுகளுக்கு எடுத்துக்காட்டுகளாகும்.

1.4.2. **எழுத்துக்கூட்டல் மரபுகள்** (Spelling Conventions)

காதிங் (Käding 1897) 11 மில்லியன் சொற்கள் கொண்ட ஜெர்மன் மொழியின் மிகப்பெரிய விரிதரவை ஜெர்மன் மொழியின் எழுத்துக்களின் வருகைகளின் மற்றும் எழுத்துக்களின் தொடர்ச்சிகளின் நிகழ்வெண்ணை கண்டுப்பிடிக்கப் பயன்படுத்தினார். இவ்விரிதரவு அதன் அளவு காரணமாக அக்காலகட்டத்தில் சிறந்ததாகக் கருதப்பட்டதுடன் சில தற்கால விரிதரவுகளுடன் ஒப்பிடுகையில் அளவு அடிப்படையில் உகந்தாகவும் இருக்கின்றது.

1.4.3. **மொழிக்கற்பித்தல்** (Language Pedagogy)

பிரைஸ் மற்றும் டிராவர் (Fries an Traver, 1940) மற்றும் போங்கர்ஸ் (Bongers, 1947) என்பவர்கள் அயல்மொழி கற்பித்தல் குறித்த ஆய்வுகளுக்கு விரிதரவைப் பயன்படுத்தினர். விரிதரவும் இரண்டாம் மொழி கற்பித்தலும் இருபதாம் நூற்றாண்டின் ஆரம்பகட்ட அரைப்பகுதியில் வலுவான தொடர்பு கொண்டிருந்தன; அயல்மொழி படிப்பவர்களின் சொற்தொகைப் பட்டியல் (Vocabulary list) விரிதரவுகளிலிருந்து எடுக்கப்பட்டன. தான்டைக் (Thorndike, 1921) மற்றும் பாமர் (Palmer, 1933) என்பவர்களின் ஆய்வுகள் அடிப்படையில் எடுக்கப்பட்ட சொல் எண்ணிக்கைகள் இரண்டாம் மொழி கற்பித்தலின் சொற்றொகைக் கட்டுப்பாட்டு இயக்கத்தின் முக்கிய இலக்காக அமைந்தது.

1.4.4. **ஒப்பீட்டு மொழியியல்** (Comparative – Linguistics)

ஒப்பீட்டு மொழியியலும் விரிதரவு பயன்படுத்தப்பட்டது. எடுத்துக்காட்டாக, ஈட்டன் (Eaton, 1940) டச்சு, ஜெர்மன், பிரெஞ்சு மற்றும் இத்தாலிய மொழிகளின் சொற்பொருள்களின் நிகழ்வெண்களை ஒப்பீட்டு செய்து ஆய்வு செய்தார். இவ்வாய்வு இன்றைய காலகட்டத்திலும் மிகச் சிறந்ததாகக்

கருதப்படுகின்றது. 1990-களின் ஆரம்பத்தில்தான் விரிதரவுகள் மீண்டும் இம்மாதிரியான தகவல்களைப் பெறுவதற்காகப் பயன்படுத்த வேண்டி உருவாக்கப்பட்டன

1.4.5. தொடரியலும் பொருண்மையியலும் (Syntax and Semantics)

ஈட்டன்-ஆல் பயன்படுத்தப்பட்ட பொருண்மையியல் நிகழ்வெண் பட்டியல்கள் ஒருமொழிய வர்ணனையில் (monolingual description) ஆர்வம் காட்டிய பிற ஆய்வாளர்களாலும் பயன்படுத்தப்பட்டது. லார்ஜ் (Lorge, 1949) இதற்கு எடுத்துக்காட்டாகும். தொடரியலும் இவ்வாறு பரிசோதிக்கப்பட்டது. ∴ப்ரைசின் (Fries, 1952) விரிதரவு அடிப்படையிலான ஆங்கில வருணனை இலக்கணம் இதற்கு ஆரம்பகால எடுத்துக்காட்டாகும். குர்க் மற்றும் பிறரின் (Quirk et al) A Comprchensive Grammar of the English Language முப்பது ஆண்டுகளுக்குப் பிந்தியதாகும். இவ்வகை ஆய்வுகள் ஆங்கிலத்தோடு நின்றுவிடவில்லை. கோகன்ஹெம் மற்றும் பிறர் (Gogenheim et al, 1956) உயர்ந்த நிகழ்வெண் உள்ள சொல் தேர்வுகளையும் இலக்கண தேர்வுகளையும் விளக்க 275 தகவலாளிகளிடமிருந்து எழுத்துப்பெயர்ப்பு செய்யப்பட்ட பேச்சுப் பிரெஞ்சு மொழியின் விரிதரவை பயன்படுத்தினர்.

இவ்வெடுத்துக்காட்டுகளிலிருந்து நீண்ட காலமாக மொழியியலில் அடிப்படை விரிதரவு நெறிமுறை விரிவாகப் பயன்படுத்தப்பட்டது என்பது தெரிய வருகிறது.

1.5. தொடக்க கால விரிதரவு குறித்த சாம்ஸ்கியின் கருத்து

சாம்ஸ்கி 1957-65-ல் மொழியியல் ஆய்வின் திசையை அனுபவவாதத்திலிருந்து (empiricism) பகுத்தறிவு வாதத்திற்கு (rationalism) குறைந்த கால கட்டத்திற்குள் மாற்றினார். இங்கு கேள்வி என்னவென்றால் அனுபவாவாதத்திற்கும் பகுத்தறிவு வாதத்திற்கும் வேறுபாடு இருக்கிறதா? இது மொழியியலுக்கு மட்டும் தனித்தன்மையானதா? இயற்கையாக நிகழும் கண்டறிதல் மீதான தீர்மானம் நம்பகமானதா? அல்லது செயற்கையாக ஊக்குவிக்கப்பட்ட உற்றுநோக்கல் நம்பகமானதா? என்ற அடிப்படையான தீர்மானத்திற்கு இது வழிவகுக்கின்றது. பகுத்தறிவாதத்தின் கோட்பாடு செயற்கையான நடத்தையின் தரவு மற்றும் உணர்வு பூர்வமான உள்முகத் தேடல் (Introspective) தீர்மானங்கள் அடிப்படையிலானது. இது ஒரு மொழியின் தாய்மொழி பேசுபவர் அம்மொழியைப் பிரதிபலிப்பதாகவும் அப்பிரதிபலிப்புகளின் அடிப்படையில் கோட்பாட்டு அடிப்படையிலான கூற்றுகளைச் செய்வதாகவும் இருக்கலாம். பகுத்தறிவு வாதத்தாரின் கோட்பாடுகள் மொழியியலின் நேர்வில் மனக் கோட்பாட்டின் வளர்ச்சியின் அடிப்படையில் அமைந்தது. இவை புலனறிவு சாத்தியத்தை அடிப்படை இலக்காகக்

கொண்டது. இதன் நோக்கம் மனித மொழியாய்வின் வெளிவிளைவுகளை வெளிப்படுத்து மட்டுமல்லாமல் உண்மையில் எவ்வாறு இவ்வாய்வு செய்யப்படுகிறது என்பதை உருப்படுத்தம் செய்கின்றது என்று கூறுகின்ற ஒரு மொழிக் கோட்பாட்டை உருவாக்குவதாகும்.

மாறாக, மொழியின் அனுபவவாத அணுகுமுறை இயற்கையாகவே நிகழும் தரவின் உற்றுநோக்கலால் செய்யப்படுகின்றது. இந்நேர்வில் ஒரு மொழியின் ஒரு வாக்கியம் சரியானதா என்று அறிய அம்மொழியின் தரவு ஆயப்பட்டு தீர்மானிக்கப்படும். இந்த அனுபவவாத-பகுத்தறிவாத பாகுபாடு கோட்பாட்டை கூற பயன்படுத்தப்படும் தரவின் இயல்பின் அடிப்படையிலானது. இந்த இரண்டு அணுகுமுறைகளுக்கும் அனுகூலங்களும் அனுகூலமின்மைகளுக்கும் உண்டு. சாம்ஸ்கி மொழியியல் ஆய்வின் நோக்கத்தை மொழியின் அருவத்தன்மையான விளக்கங்களிலிருந்து மொழியின் உளவியல் உண்மையையும் புலனறிவுசார் சாத்தியமான மாதிரிகளையும் பிரதிபலிக்கும் கோட்பாடுகளுக்கு மாற்றினார். சாம்ஸ்கி இதன்படி மொழியாய்வில் பெருந்தரவை சான்றின் ஒரு மூலமாகக் கருதுவதை மதிப்பில்லாமல் ஆக்கினார். சாம்ஸ்கி மொழியியலாரின் மாதிரி செயல்திறனை விட அறிதிறனைதான் அடிப்படையில்தான் தங்கள் மாதிரியை அமைக்க முயல வேண்டும் என்ற காரணத்தால் பெருந்தரவு மொழியியலாருக்குப் பயனுள்ள கருவியாக இருக்கவியலாது என்று கூறினார்.

அறிதிறன் ஒரு மொழியின் உள்ளுறை செய்யப்பட்ட அறிவாகும். ஆனால் செயல்திறன் மொழி அறிதிறனின் புறச் சான்றாகும். சாம்ஸ்கி மொழியியலார் செயல்திறனைவிட அறிதிறனைத் தான் மாதிரியாக்கம் செய்ய முயன்றனர் என்ற வாதிட்டார். செயல்திறன் அறிதிறனைப் பிரதிபலிக்காது என்று வாதிடப்பட்டது. செயல்திறன் நமது அறிதிறனிலிருந்து விலகிய காரணிகளால் பாதிக்கப்டலாம். ஒரு பெருந்தரவு அதன் இயல்பு அடிப்படையில் வெளிப்படுத்தப்பட்ட கூற்றுகள் ஆகும். அது செயல்திறனின் தரவாகும். இதன்காரணமாக அது மொழி அறிதிறனை மாதிரிப்படுத்த மோசமான வழிகாட்டியாகும். சாம்ஸ்கி அனுபவவாதத்திலிருந்து பகுத்தறிவாதத்திற்கு மாறத் தூண்டினார். மொழியை உற்றுநோக்கல் அடிப்படையில் விளக்காமல் உள்முகத் தேடல் (சிந்தனை) அடிப்படையில் விளக்க வேண்டும் என்றார்.

1.6. விரிதரவு மொழியியலுக்கான வாதங்கள்

விரிதரவு அடிப்படையிலான மொழியியல் ஆய்வுகளுக்கு எதிரான வாதங்கள் எல்லா விரிதரவு அடிப்படையிலான ஆய்வுகளையும் தடைசெய்து விடவில்லை. ஒலியியலில் இயல்பாகப் பெறப்பட்ட தரவுகள் சான்றுகளுக்கான ஆதிக்க மூலமாக இருந்துவந்தது. மொழிப்பேறு ஆய்வுக் களத்திலும்

இயல்பாகக் கிடைக்கும் தரவுகள் ஆதிக்கம் செய்தன. குழந்தைகளின் மொழிப்பேறை ஆய்கின்ற மொழியிலார்களுக்கும் உளவியலார்களுக்கும் உள்முகத்தேடல் தீர்மானங்கள் (inrospective judgements) கிடைக்கவில்லை. சாம்ஸ்கியே (1964) செயல்திறன் தரவுகளைச் சான்றின் மூலமாகக் கருதாது தள்ளி வைத்தல் மொழிப்பேறு ஆய்வுகளுக்குப் பொருந்தாது என்று கூறினார். பொதுவாகக் கூறினால் 1960களிலும் 1970களிலும் விரிதரவு நெறிமுறை நடைமுறை படுத்தப்பட்டது.

இயல்பாகக் கிடைக்கும் தரவு அதைப் பரிசோதனை செய்ய விரும்புகின்ற எல்லோராலும் உற்றுநோக்கப்படுவதற்கும் சரிபார்க்கப்படுவதற்கும் உட்பட இயலும் என்ற நிலையில் பயனுள்ளது. பேசுபவர் ஒரு உள்முகத்தேடல் தீர்மானத்தைச் செய்கையில் அது சரியானது என்று எவ்வாறு கூற இயலும்? அவர் ஒரு வாக்கியத்தைக் கூறும்போது நாம் அதை உற்றுநோக்கி பதிவு செய்யவியலும். ஆனால் அவர் தன் கருத்தை ஒரு சிந்தனைச் செயற்பாங்கால் வெளியிட்டால் நாம் என்ன செய்யவியலும்? அது உற்றுநோக்கவியலாததாக இருக்கும். பதிவு செய்யப்பட்ட வாக்கியத்தால் நாம் மக்களின் கருத்தைப் பெறவியலும். தரவு எல்லோராலும் உற்றுநோக்கத்தக்கதும் விமரிசிக்கத்தக்கதும் ஆகும். மக்கள் கருத்து மற்றும் அதற்கு மாறான தனிநபர் கருத்து என்ற சிக்கல் மொழியியலை மட்டுமற்றி பிற துறைகளையும் பாதித்துள்ளது. விரிதரவு அடிப்படையிலான உற்றுநோக்கல்கள் உள்முகத்தேடல் அடிப்படையிலான தீர்மானங்களைவிட மெய்யாகச் சரிபார்க்க உகந்தது.

செயற்கையான தரவு செயற்கையானதாகத் தான் இருக்கும். சாம்சன் (Samson, 1992:428) உள்முகத்தேடல் மொழியியலார்களால் ஆயப்பட்ட வாக்கிய வகை விரிதரவில் கிடைக்கும் சான்றுகளின் வகையிலிருந்து மிக விலகி இருக்கின்றது என்று கூறுகின்றார்.

நிகழ்வெண் மற்றொரு செய்தியை முன்வைக்கின்றது. சில வகை மொழித் தரவுகளை விரிதரவிலிருந்து மட்டுமே சரியாகப் பெறவியலும். மனிதவினம் ஒரு கட்டுமானத்தின் அல்லது சொல்லின் நிகழ்வெண் குறித்த மிகத் தெளிவற்ற கருத்துச்சாயலைத்தான் கொண்டிருக்கின்றது. தரவின் இயல்பான உற்றுநோக்கல் மட்டும் தான் நிகழ்வெண் போன்ற பண்புக்கூறுகளுக்கான சான்றின் நம்பகமான மூலமாக இருக்கின்றது. சாம்ஸ்கி இம்மாதிரியான அளவுசார் தகலைப் (quantitative information) பெறுதல் நன்மை பயப்பதல்ல என்று இச்செய்திக்கு எதிராகக் கூறலாம். லீச் (1992) விரிதரவு விளைவுகளின் புறவயமான சரிபார்ப்புக்கு (objective verification) அனுமதிப்பதால் அறிவியல் நெறிமுறைப் பார்வையின் படி விரிதரவு கூடுதல் வலுவான நெறிமுறையாகும் என்று வாதிடுகின்றார்.

சாம்ஸ்கி (Chomsky, 1965) மேம்போக்காக 95 விழுக்காடு கூற்றுகள் இலக்கணத்தன்மையற்றது என்று வகைப்படுத்துகின்றார்; எனவே செயல்திறன்சார் தரவு அறிதிறனின் தவறான உருப்படுத்தம் என்கின்றார். இக்கூற்று மிக எளிதாக்கப்பட்ட ஒன்றாகும். லபோவ் (Labov, 1969) எல்லாச் சூழல்களிலும் கூற்றுக்களின் மிகப்பெரும்பான்மையானவை இலக்கணத்தன்மையானவை என்று காட்டுகின்றார். விரிதரவு அடிப்படையிலான ஆய்வு ஆரம்பத்தில் கருதப்பட்டதுபோல் செல்லாதது அல்ல. விரிதரவு இலக்கணத்தன்மையற்ற வாக்கியங்களின் தொகுதியாக இருக்கத் தேவையில்லை. விரிதரவு பொதுவாக இலக்கணத்தன்மையான வாக்கியங்களைக் கொண்டிருக்கும் என்று நம்புவதற்குக் காரணம் இருக்கின்றது என்று தோன்றுகின்றது. நாம் விரிதரவிலுள்ள எல்லா வாக்கியங்களும் இலக்கணத் தன்மையானவை என்று கூறவில்லை.

விரிதரவுகள் அளவுசார் தரவின் (quantative data) மிகச் சிறப்பான மூலங்களாகும். ஆனால் சாம்ஸ்கி அளவுசார் தரவு மொழியியலுக்குப் பயன்படாது என்று பதில் கூறலாம். ஆனால் அவருடைய கருத்து உண்மை நிகழ்வால் ஆதரிக்கப்படவில்லை என்று நாம் கூறலாம். விரிதரவு அல்லது சான்றின் ஒப்பிடவியலும் அளவுசார் தரவைத் தரும் சில பிற இயற்கை மூலமில்லாமல் மொழியியலார்களுக்கோ கணினி மொழியலார்களுக்கோ வலுவான ஆய்வுக் கருவிகள் இருந்திருக்காது.

சாம்ஸ்கியால் விரிதரவு பற்றி கூறப்பட்ட சில திறனாய்வுகள் பகுதி செல்லத்தக்கதாகும்; அவை இன்றைய விரிதரவுகள் பற்றி கூடுதல் மெய்யான மனப்பாங்கு வளர உதவுகின்றது. ஆனால் இத்திறனாய்வுகள் (குறைகள்) பகுதி செல்லாதது; அவை விரிரிதரவின் வலுவின்மைக்கு எதிரான வலுவின் உண்மையான கணிப்பை விலக்குகின்றது. இவ்வுற்றுநோக்கல் ஏன் சிலர் விரிதரவு அடிப்படையில் ஆய்வை மேற்கொண்டு வருகின்றனர் என்றும் ஏன் அவர்கள் ஆரம்ப கால விரிதரவு குறித்த முக்கியமான குறைகூறலை நேரிட்டாலும் அதைத் தொடர்கின்றனர் என்றும் குறிப்பிடுகின்து.

இப்போது விரிதரவு என்பது இயந்திரத்தால் படிக்கவியலும் விரிதரவு என்ற சொல்லின் ஒருபொருள் பன்மொழியாகப் பயன்படுத்தப்படுகின்றது. இது எதேட்சையானது அல்ல. கணிப்பொறி பயன்படுத்தவேண்டிய போலி உத்தி என்ற வட்டத்திற்குள் இருப்பதாக முன்னர் கருதப்பட்ட உத்திகளை அனுமதிக்கின்து. கணிப்பொறி அதன் தரவைத் தேடும், மீட்கும் மற்றும் கணிக்கும் திறமை காரணமாக இதை அனுமதிக்கின்றது.

பகுதி 2

விரிதரவு மொழியலின் இன்றைய நிலை

2.0. அறிமுகம்

மின் விரிதரவின் அறிமுகம் 1960களில் ஏற்பட்டது. நாற்பது ஆண்டுகளுக்குள் தனிநபரின் ஆர்வத்தாலோ கல்வி மற்றும் ஆய்வு நிறுவனங்களின் ஊளக்கத்தாலோ உலகின் பல பாகங்களிலும் பல மொழி விரிதரவுகள் திட்டமிடப்பட்டு உருவாக்கப்பட்டுள்ளன. நாம் அடிப்படையான பண்புக்கூறுகள் அடிப்படையில் விரிதரவுகளை இரண்டு வகைகளாகப் பகுக்கலாம்: அவை எழுத்து விரிதரவு மற்றும் பேச்சு விரிதரவு. விரிதரவுகளின் சில எடுத்துக்காட்டுகள் இங்கு தரப்பட்டுள்ளன. ஒவ்வொன்றும் அவற்றின் தகவல் மற்றும் பொருளடக்கம் அடிப்படையில் விளக்கப்பட்டுள்ளன.

2.1. எழுத்து விரிதரவு

2.1.1. பிரவுன் விரிதரவு

முதல் மின் விரிதரவு 1961-இல் பிரவுன் பல்கலைக் கழகத்தில் நெல்சன் பிரான்சிஸ் ஹென்றி குச்சேரா (W. Nelson Francis and Henry Kuchera) என்பவர்களால் உருவாக்கப்பட்ட பிரவுன் விரிதரவாகும் (Brown Corpus (The standard sample of present day American English)). இது 1961-ஆம் ஆண்டு (நாள் காட்டி ஆண்டு) யுஎஸ்ஏ-இல் (USA) அச்சடிக்கப்பட்ட ஆங்கில உரை நடையின் தொடர்ச்சியான பனுவலின் 1014312 சொற்களைக் கொண்டது. இது 2000+ சொற்களைக் கொண்ட 500 மாதிரிகளாகப் (sample) பகுக்கப்பட்டுள்ளது. ஒவ்வொரு மாதிரியும் ஒரு வாக்கியத்தின் தொடக்கத்தில் தொடங்குகிறது; ஆனால் ஒரு பத்தியின் அல்லது ஒரு பெரிய பகுதியின் தொடக்கத்தில் தொடங்குகிறது என்பது தேவை இல்லை; ஒவ்வொன்றும் 2000 சொற்களுக்குப் பின் இறுதியுறும் முதல் வாக்கியத்தில் முடிகிறது. ஒவ்வொரு மாதிரியும் உரைநடையின் நடைகள் மற்றும் வகைகளின் விரிந்த பரப்பெல்லையை பிரதிநித்துவம் செய்யும். செய்யுள்களும் நாடகங்களில் உள்ள

உரைகளும் எழுத்து கருத்தாடல்களுக்கு அப்பாற்பட்ட பேச்சு கருத்தாடலின் கற்பனையான பொழுது போக்காக இருப்பதன் காரணமாக அவைகள் விலக்கி வைக்கப்படுகிறது. கதைகள் உட்படுத்தப்பட்டுள்ளன; ஆனால் 50 விழுக்காடுக்கு மேல் உரையாடல் உள்ள மாதிரிகள் எடுத்துக்கொள்ளப்படமாட்டாது. மாதிரிகள் அவற்றின் பிரதிநித்துவப் பண்பிற்கேயன்றி வேறு எந்த அகவயமாகத் தீர்மானிக்கப்பட்ட சிறப்புக்காகவும் தெரிந்தெடுதக்கப்படவில்லை.

தேர்வு செயல் முறைகள் முன்று நிலைகளைக் கொண்டது:

1. தொடக்க நிலை அகவய வகைப்பாடு
2. ஒவ்வொரு வகைப்பாட்டிலும் எவ்வளவு மாதிரிகள் பயன்படுத்தப்படும் என்பதன் தீர்மானம்
3. ஒவ்வொரு வகைப்பாட்டிற்குள்ளும் மெய்யான மாதிரிகளின் வரைமுறையற்ற தேர்வு

விரிதரவுக்கு இரண்டு முக்கிய பகுதிகள் இருக்கின்றன:

1. தகவல் உரைநடையின் உரைகள்
2. கற்பனை உரைநடையின் உரைகள்

தகவல் உரைகள் பின்வருவனவற்றை உள்ளடக்கும்:

- பத்திரிக்கை: அறிக்கை (அரசியல், விளையாட்டு, சமுதாயம், வட்டாரச் செய்திகள், பொருளாதாரம், பண்பாடு)
- பத்திரிக்கை: இதழாசிரியர் கருத்து (நிறுவனம்சார்ந்தவை, பதிப்பாசிரியருக்கு தனிநபர் கடிதங்கள்)
- பத்திரிக்கை: திறனாய்வு (சினிமா, புத்தகம், இசை, நடனம், போன்றவை)
- மதம் (புத்தகங்கள், கால இதழ்கள், சமயக் கட்டுரைகள்)
- திறமைகளும் பொழுதுபோக்குகளும் (புத்தகங்கள், கால இதழ்கள்)
- புகழ் வாய்ந்த மரபுச் செய்திகள் (புத்தகங்கள்,கால இதழ்கள்)
- கடிதங்கள், வாழ்க்கை வரலாறு, நினைவுகள் (புத்தகங்கள், கால இதழ்கள்)
- பலதரப்பட்டவை (அரசாங்க ஆவணங்கள், அடித்தள அறிக்கைகள், தொழில் நிறுவன அறிக்கைகள், கல்லூரி பெயர் பட்டியல், தொழில் நிறுவனம், வீடு)

- கற்றல் மற்றும் அறிவியல் எழுத்துரைகள் (இயற்கை அறிவியல், மருத்துவம், கணித்தொழில் நுட்பம், பொறியியல்)
- சமூக மற்றும் நடத்தை அறிவியல்கள் (அரசியல் அறிவியல், சட்டம், கல்வி)
- மனிதவியலிலிருந்து உரைகள்.

கற்பனை உரைநடையின் உரைகள் பின்வருவனவற்றை உள்ளடக்கும்:

- பொதுக் கதைகள் (புதினம், சிறுகதைகள்)
- அறிவியல் கதைகள் (புதினங்கள், சிறுகதைகள்)
- துணிக்கரச் செயல் பற்றிய மற்றும் மேற்கத்திய கதைகள் (புதினங்கள், சிறுகதைகள்)
- காதல் கதைகள் (புதினங்கள், சிறுகதைகள்)
- நகைச்சுவை (புதினங்கள், சிறுகதைகள்)

2.1.2. லோப் விரிதரவு

லங்காஸ்டர் ஓசலோ பெர்கன் விரிதரவு (Lancaster Oslo/Bergen (LOB) என்பது பிரிட்டிஷ் ஆங்கிலத்தின் ஒரு மில்லியன் சொல் சேகரிப்பாகும். இது லங்காஸ்டர் பல்கலைக் கழகத்தைச் சார்ந்த ஜியாபிரே லீச் (Geofferey Leech) மற்றும் ஓசலோ பல்கலைக் கழகத்தைச் சார்ந்த ஸ்டிக் ஜோகன்சன் (Stig Johansson) என்போர் பெர்கனில் உள்ள மனிதவியலுக்கான நார்வேஜியன் கணிப்பு மையத்துடன் இணைந்து தொகுப்பட்டதாகும். லோப் யுகே-இல் (UK) 1961-ஆம் வருடத்திற்குள் வெளியிடப்பட்ட மூலப்பொருட்களைக் கொண்டது. இது பிரவுண் விரிதரவைப் போல் உரை மாதிரிகளைக் கொண்டது. லோப் விரிதரவின் திட்ட நெறிமுறை யுஎஸ்ஏ-இலும் யுகே-இலும் பயன்படுத்தப்படும் ஆங்கிலத்தின் இரு வகைகளுக்கிடையே உள்ள எதிர்கால ஒப்பீட்டிற்கு வேண்டி பிரவுண் விரிதரவுன் ஒற்றுமை கொண்டுள்ளது.

2.1.3. ஆங்கிலத்தில் ஆஸ்டிரேலியன் விரிதரவு

ஆங்கிலத்தில் ஆஸ்டிரேலிய விரிதரவு (Australian Corpus of English (ACE)) என்பது ஆஸ்டிரேலியாவிலுள்ள மாக்குயர் பல்கலைக் கழகத்தில் (Macquaire University) மொழியியல் துறையில் 1986-1990 என்ற கால கட்டத்தில் தொகுக்கப்பட்டது. ஏசியி 1986-க்குள் வெளியிடப்பட்ட

மூலப்பொருட்களைக் கொண்டிருக்கின்றது. ஏசியி பல்வேறுபட்ட மொழியியல் ஆய்வுகளில் உதவுவதற்காக திட்டமிடப்பட்ட ஆஸ்டிரேலியாவில் முதலில் தொகுக்கப்பட்ட பலபடித்தான விரிதரவாகும். இது தற்கால ஆஸ்டிரேலிய ஆங்கிலத்தின் திறனான மாதிரியாகும். இது ஆஸ்டிரேலியாவில் கூடுதல் சிறப்புப்பண்புள்ள ஒருபடித்தான விரிதரவுகளுடன் ஒப்பீட்டுப் பார்ப்பதற்கு வேண்டி நோக்கீட்டு விரிதரவாகச் செய்யப்படுகின்றது. இது பிரவுண் விரிதரவிலும் லோப் விரிதரவிலும் பிரதிநிதித்துவம் தரப்பட்டுள்ள இனங்களின் சரிசமத்திற்குப் பொருத்துவதை நோக்கமாகக் கொண்டது. எனவே இது ஏறக்குறைய ஒவ்வொரு வகைப்பாட்டிலும் 2000 சொல் மாதிரிகளின் நிகரான குழுமத்தை உருவாக்குகிறது. துணை இனங்களையும் பாடப்பரப்புகளையும் பொருத்துவது தேவை என்பதால் ஒவ்வொரு விரிதரவு வகைப்பாட்டிற்குள்ளும் மாதிரிச் செயல் முறைகள் பெரும்பாலும் திட்டமிடப்பட்டுள்ளது (வரைமுறையற்றதல்ல).

2.1.4. எழுதப்பட்ட நியூசிலாந்து ஆங்கிலத்தின் வெல்லிங்டன் விரிதரவு (Wellington Corpus of Written Newzealand English)

எழுதப்பட்ட நியூசிலாந்து ஆங்கிலத்தின் வெல்லிங்டன் விரிதரவு (Wellington Corpus of Written Newzealand English (WCWNEZE)) என்பது 1982 கால கட்டத்தில் வெல்லிங்டன் விக்டோரியா பல்கலைக் கழகத்தில் (Wellington Victoriya University) மொழியியல் துறையில் உருவாக்கப்பட்டது. நியூசிலாந்து ஆங்கில விரிதரவின் நோக்கம் பிரவுன் விரிதரவு, லோப் விரிதரவு, ஏசியி விரிதரவு என்பதுடன் நேரடியாக ஒப்பீடு செய்வதை அனுமதிக்கும் எழுத்து நீயூசிலாந்து ஆங்கிலத்தின் கணினியாக்கம் செய்யப்பட்ட மாதிரியைத் தருகிறது. முக்கிய உரை வகைப்பாடுகள் எவ்வளவு இயலுமோ அவ்வளவு லோப் விரிதரவுடன் பொருந்தும் படி வரிசைப்படுத்தப்பட்டுள்ளன. கற்பனை உரை நடையில் குறிப்பாகக் கதைகளில் தேவை கருதிச் சில மாற்றங்கள் செய்யப்பட்டுள்ளன.

2.1.5. இந்திய ஆங்கிலத்தில் கோலாப்பூர் விரிதரவு

இந்திய ஆங்கிலத்தில் கோலாப்பூர் விரிதரவு (Kolhapur Corpus of Indian English (KCIE) எஸ்.வி சாஸ்த்திரியும் அவரைச் சார்ந்தவர்களும் இணைந்து கோலாப்பூரில் உள்ள சிவாஜி பல்கலைக் கழகத்தில் 1988-ஆம் ஆண்டில் வெளியிடப்பட்ட விசயங்களிலிருந்து எடுக்கப்பட்ட இந்திய ஆங்கிலத்தின் கிட்டத்தட்ட ஒரு மில்லியன் சொற்களைக் கொண்டது. இவ்விரிதரவு அமெரிக்க, பிரிட்டிஷ் மற்றும் இந்திய ஆங்கிலத்தின் ஒப்பீட்டு ஆய்வுகளுக்கு வேண்டிய மூல சாதனங்களாகப் பயன்படுகின்றது. இது இந்திய ஆங்கிலத்தின் விரிந்த வர்ணனைக்குக் கொண்டுச் செல்லும் என்று

எதிர்பார்க்கப்படுகிறது. லோப் மற்றும் பிரவுன் விரிதரவுகளை உருவாக்க பயன்படுத்தப்பட்ட திட்டத்தைப் பின்பற்றி கோலப்பூர் விரிதரவுக்கு 15 பாட வகைப்பாடுகளிலிருந்து உரைகள் தேர்ந்தெடுக்கப்பட்டுள்ளன.

2.1.6. ∴ப்லோப் விரிதரவு

∴ப்லோப் விரிதரவு (Freiburg LOB Corpus for British English (FLOB)) பிரவுன் மற்றும் லோப் விரிதரவுகளுடன் பொருத்தும் ஒரு குழும விரிதரவுகளைத் தொகுக்கும் முயற்சியின் விளைவாகும். இது 1990-களின் தொடக்க நிலை மொழியைப் பிரதிநிதித்துவம் செய்யும் என்ற நிலையில் வேறுபடும். இவ்வாய்வுத் திட்டம் 1990 ஏப்பிரலில் ஜெர்மனியில் கிருஸ்டியன் மேயிரின் கீழ் தொடங்கப்பட்டது. இத்திட்டத்தின் நோக்கம் மொழியில் தொடர்ந்து நடைபெறும் மாற்றத்தை ஆயவேண்டி அனுபவவாத அடிப்படையை மொழியியலாலருக்குத் தருவதாகும்.

இது மொழியியலாளர்களுக்குப் பின்வரும் செயல்களில் உதவுகிறது: (1) இன்றைய ஆங்கிலத்தில் மொழிசார் மாற்றத்தின் மீதான இன்றைய கருதுகோள்களைப் பரிசோதிக்க, (2) சொல் நிகழ்வெண்களின் முறையான ஒப்பீட்டுகளின் வழி இலக்கியத்தில் இதுவரை கண்டுபிடிக்கப்படாத மாற்றங்களைக் கண்டுப்பிடிக்க, (3) ஒருகால இடவழக்குகளின் (British மற்றும் America) தொடர்ச்சியான மாற்றம் மற்றும் நடையியல் வேறுபாடு இவற்றிற்கு இடையிலான சார்பையும் மெய்யான இருகால வளர்ச்சிகளையும் ஆயும் முக்கியமான நெறிமுறைச் சிக்கல்களில் ஒன்றை முறையாக கையாளுவது. இப்புதிய பிரிட்டிஷ் மற்றும் அமெரிக்க விரிதரவுகளின் கூடுதல் அனுகூலம் என்னவென்றால் அவை இந்திய, ஆஸ்திரேலிய மற்றும் நியூசிலாந்து விரிதரவுகளுடன் (பிந்தைய 1980-களில் உள்ள மொழிப்பயன்பாட்டை பிரதிநித்துவம் செய்யும் மாதிரிகள்) ஒப்பிட மூல லோப் மற்றும் பிரவுண் விரிதரவுகளைவிடக் கூடுதல் பொருத்தமான தரவு மையங்களைத் தருகின்றன.

2.1.6. பிரிட்டிஷ் தேசிய விரிதரவு

பிரிட்டிஷ் தேசிய விரிதரவு (British National Corpus (BNC)) என்பது மூலங்களின் விரிந்த பரப்பெல்லையிலிருந்து பெறப்பட்டு எழுதப்பட்ட மற்றும் பேசப்பட்ட மாதிரிகளின் சேகரிப்பைக் கொண்ட தற்கால ஆங்கிலத்தின் மிகப்பெரிய (100 மில்லியன் சொற்களுக்கு மேல்) விரிதரவாகும். இது பேசப்படும் மற்றும் எழுதப்படும் தற்கால பிரிட்டஷ் (current British) ஆங்கிலத்தின் பரந்த பகுதியைப்

பிரதிநிதித்துவம் செய்வதற்கு வேண்டி திட்டமிடப்பட்டுள்ளது. இது சில நிறுவனங்கள் அடங்கிய தொழில்/கல்வி குழுவால் உருவாக்கப்பட்டது. விரிதரவை உருவாக்கும் வேலை 1991-இல் தொடங்கி 1994-இல் முடிக்கத் திட்டமிடப்பட்டது. ஐரோப்பிய ஆய்வாளர்களுக்கு பிஎன்சியின் முதல் பொதுவான வெளியீடு பெப்ரவரி 1995-இல் அறிவிக்கப்பட்டது.

பிஎன்சி தற்கால பிரிட்டிஷ் ஆங்கிலத்தின் விரிந்த பரப்பெல்லையை பிரதிநிதித்வம் செய்ய வேண்டி வடிவமைக்கப்பட்டுள்ளது. எழுத்துப் பகுதி (90%) வட்டார மற்றும் தேசீய செய்தித்தாள்கள், எல்லா வயதினர்களுக்கும் ஆர்முடைவர்களுக்கும் வேண்டி வெளிவரும் சிறப்பான காலவிதழ்கள் மற்றும் பத்திரிகைகள், கல்விசார் புத்தகங்கள் மற்றும் புகழ்வாய்ந்த கதைப் புத்தகங்கள், வெளியிடப்பட்ட மற்றும் வெளியிடப்படாத கடிதங்கள் மற்றும் அறிக்கைகள், பள்ளி மற்றும் பல்கலைக்கழகக் கட்டுரைகள் மற்றும் பிற பல உரை வகைகளை உள்ளடக்கும். பேச்சுப்பகுதி (10%) பல வயதினர்கள், வட்டாரங்கள், சமூக வகுப்புகள் இவற்றிலிந்து பெறப்பட்ட எழுதப்படாத முறைசாரா உரையாடல்கள் மற்றும் முறையான வணிக மற்றும் அரசாங்க கூட்டங்களிலிருந்து வானொலி மற்றும் தெலைபேசி வரையிலான பரப்பெல்லையைக் கொண்ட சூழல்களின் எல்லா வகைகளிலிருந்தும் சேகரிக்கப்பட்ட பேச்சு மாதிரிகள் என்பனவற்றை உள்ளடக்கும்.

இப்பொதுத்தன்மை பிஎன்சி விரிதரவை அகராதியியல், பொருண்மையியல், உருபனியல், மொழிப் பகுப்பாய்வு, செயற்கை அறிவுநுட்பம், பேச்சு அறிதல் மற்றும் கூட்டிணைப்பு, இலக்கிய ஆய்வுகள் மற்றும் முக்கிய மொழியியலின் எல்லாப் பரப்புகளையும் உள்ளடக்கும் ஆய்வு நோக்கங்களின் விரிந்த வகைக்குப் பயனுள்ளதாகச் செய்கின்றது.

பிஎன்சி பேச்சு உரையாடல் மற்றும் ஒருவர் உரையாடல் இவற்றிலிருந்து ஒலிபெயர்ப்பு செய்யப்பட்ட 863 உரைகளை உள்ளடக்கிய 4124 உரைகளைக் கொண்ட 100106008 சொற்களை உள்ளடக்கியது. ஒவ்வொரு பனுவலும் மேலும் எழுத்துவடிவ வாக்கிய அலகுகளாகப் பிரிக்கப்பட்டுள்ளது; இதற்குள் ஒவ்வொரு சொல்லுக்கும் சொல் வகுப்பு (சொல்வகைப்பாடு) குறியம் தரப்பட்டுள்ளது. பிரித்தலும் சொல் வகைப்படுத்தலும் CLAWS சொல்வகைப்பாட்டு அடையாளப்படுத்தியால் தானியக்கமாகச் செய்யப்படுகின்றது. பிஎன்சி விரிதரவில் பயன்படுத்தப்படும் வகைப்பாட்டுத் திட்டம் 65 சொல் வகைப்பாடுகளை வேறுபடுத்துகின்றது.

2.1.7. அமெரிக்க தேசிய விரிதரவு

அமேரிக்க தேசிய விரிதரவு (American National Corpus (ANC) பிஎன்சி விரிதரவுடன் ஒப்பிடத்தக்க அமெரிக்க ஆங்கிலத்தை உட்படுத்திய விரிதரவின் உருவாக்கத்திற்கு உதவியது. இதன் நோக்கம் பிஎன்சி விரிதரவுப் பொதுவினங்களுடன் ஒப்பிடத்தக்கக் குறைந்தது 100 மில்லியன் சொற்களைப் பெறுவதாகும். இது மொழித் தொழில் நுட்ப ஆய்வுக்கு உதவுவதுடன் எல்லா மட்டங்களிலும் மொழிப் பகுப்பாய்வுக்கும் கல்விக்கும் பயன்பட வேண்டி செழுமையான தேசிய மூலத்தைத் தருகின்றது. ஏஎன்சி அமெரிக்க ஆங்கில அகராதிகளின் வெளியீட்டாளர்கள் மற்றும் மொழி ஆய்வில் ஆர்வமுள்ள நிறுவனங்கள் இவற்றின் கூட்டிணைப்பின் (Consortium) பங்களிப்பு வழி உருவாக்கப்பட்டது. மொழித் தரவுக் கூட்டிணைப்பும் (Linguistics Data Consortium (LDC)) சாதன மற்றும் பொருளாதார ஆதரவு தந்தது. ஏஎன்சி விதரவின் 10 மில்லியன் சொற்களின் முதல் வெளியீட்டில் உள்ளடக்கப்பட்ட உரைகள் முதலில் பெறப்பட்டவையாகும். எனவே விரிதரவு சரிநிகரானதல்ல: பேச்சுப் பகுதி 3224388 சொற்களையும் எழுத்துப்பகுதி 8283828 சொற்களைக் கொண்டிருக்கின்றது.

2.1.8. ஆங்கில வங்கி

ஆங்கில மொழி வங்கி (Bank of English) சொற்கள், இலக்கணம் மற்றும் பயன்பாடு (usage) என்பனவற்றின் பகுப்பாய்விற்கு வேண்டி வடிவமைக்கப்பட்ட தற்கால ஆங்கில மொழியின் மாதிரிகளின் சேகரிப்பாகும். ஆங்கில மொழி வங்கி 1991-இல் COBUILD (Harper Collins Publishers) மற்றும் யுகேயில் உள்ள பிர்மிங்காம் பல்கலைக்கழகம் இவற்றால் 1991-இல் வெளியிடப்பட்டது. 2002 சனவரியில் 450 மில்லியன் சொற்களைக் கொண்ட புதிய விரிதரவு வெளியிடப்பட்டது. இது புதிய சாதனங்களின் நிரந்ததர சேர்ப்பினால் தொடர்ந்து வளர்கின்றது. சேகரிப்பு எழுதப்பட்ட மற்றும் பேச்சின் வேறுபட்ட வகைகளின் விரிந்த பரப்பெல்லையைக் கொண்டது. இது நூற்றுக்கணக்கான வேறுபட்ட மூலங்களிலிருந்து ஆங்கில மொழியின் மாதிரிகளைக் கொண்டிருக்கின்றது. எழுதப்பட்ட உரைகள் செய்தித்தாள்கள், இதழ்கள் (magazines), கதைப் புத்தகங்கள் மற்றும் கதையல்லாப் புத்தகங்கள், சிற்றேடுகள் (brochers), துண்டுப்பத்திரிக்கைகள் (leaflets), அறிக்கைகள், கடிதங்கள் மற்றும் பிறவற்றிலிருந்து எடுக்கப்பட்டவைகளாகும். பேச்சுச் சொற்கள் அன்றாட சாதாரண உரையாடல்கள், வானொலி ஒலிபரப்புகள், கூட்டங்கள், நேர்காணல்கள் மற்றும் விவாதங்கள் போன்றவற்றால் பிரதிநிதித்துவம் செய்யப்பட்டுள்ளது.

2.1.9. இந்திய மொழிகளின் MIT விரிதரவு

இந்திய மொழிகளுக்கு விரிதரவு உருவாக்குவது 1991-இல் தொடங்கப்பட்டது. இது இந்திய அரசின் மின்னனுத் துறை எல்லா இந்திய மொழிகளுக்கும் இயந்திரத்தால் படிக்கவியலும் உரைகளின் விரிதரவுகளை உருவாக்க வேண்டி இந்திய மொழிகளுக்கு தொழில் நுட்ப வளர்ச்சி (Technological Development for Indain Languages) என்ற திட்டத்தைத் தொடங்கியது. இதே நேரத்தில் மொழி ஆய்வுக்கும் (சொல்வகைப்பாடு அடையாளப்படுத்தி, உரை குறியாக்கி, புள்ளியல் எண்ணி, எழுத்துத் திருத்தி, உருபனியல் பகுப்பாய்வி போன்றவற்றை உருவாக்க) மற்றும் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு எந்திர உதவிசார் மொழிபெயர்ப்புக்கு (machine-aided translation) வேண்டி கருவிகளை வடிவமைத்தல் என்ற செயல்பாடுகளுக்கு மென்பொருள்களை உருவாக்கவும் ஊக்கம் தரப்பட்டது. டெல்கியில் உள்ள இந்தியத் தொழில் நுட்ப நிறுவனம் (Indian Institute of Technolgy) இந்திய ஆங்கிலம், இந்தி மற்றும் பஞ்சாபி ஆகிய மொழிகளுக்கு விரிதரவுகளை உருவாக்கியது; மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனம் தமிழ், தெலுங்கு, கன்னடா மற்றும் மலையாள மொழிகளுக்கு விதரவுகளை உருவாக்கியது; பூனேயில் உள்ள டெக்கான் கல்லூரி மராத்தி மற்றும் குஜராத்தி ஆகிய மொழிகளுக்கு விரிதரவுகளை உருவாக்கியது; புபனேஸ்வரிலுள்ள பயன்பாட்டு மொழி அறிவியல்களின் இந்திய நிறுவனம் ஒரியா, வங்காளம், அசாமீஸ் ஆகிய மொழிகளுக்கு விரிதரவுகளை உருவாக்கியது; வாரணாசியிலுள்ள சம்பூர்ணநந்தா சமஸ்கிருதப் பல்கலைக்கழகம் சமஸ்கிருத மொழிக்கு விரிதரவு உருவாக்கியது; அலிகார் முஸ்லீம் பல்கலைக்கழகம் உருது, சிந்தி மற்றும் காஷ்மீரி ஆகிய மொழிகளுக்கு விரிதரவுகளை உருவாக்கியது.

கான்பூரிலுள்ள இந்தியத் தொழில் நுட்ப நிறுவனம் மொழி பகுப்பாய்விற்கும் இயந்திர மொழிபெயர்ப்புக்கும் கருவிகளையும் மென்பொருள்களையும் வடிவமைக்கும் பொறுப்பை ஏற்றது. மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனம் எல்லா மொழிகளின் முழு விரிதரவு தரவுமையத்தைச் சேகரித்துப் பாதுகாக்கும் பொறுப்பை ஏற்றது. 1995-இல் இத்திட்டம் முடிவுற்றபோது மின்னணு நிறுவனம் இது நீண்டகால ஆய்வையும் பெரிய முதலீட்டையும் வேண்டும் என்று உணர்ந்து திட்டத்தைத் தொடராமல் நிறுத்தியது. இருப்பினும் தற்போது இந்திய அரசின் MIT புதிய வெளிச்சத்திலும் பார்வையிலும் இச்செயல்பாட்டை மீண்டும் நடைமுறைப்படுத்தியுள்ளது.

2.2. பேச்சுத் தரவு

2.2.1. பேச்சு ஆங்கிலத்தின் லண்டன்-லண்ட் விரிதரவு

லண்டன-லண்ட் விரிதரவு (London-Lund Copurs of Spoken English (LLC)) இரு திட்டங்களின் விளைவுகளைக் உட்படுத்தியது: (1) 1961-இல் லண்டன் பல்கலைக்கழகக் கல்லூரியில் ராண்டோல்ஃப் குயிர்கால் செயல்படுத்தப்பட்ட ஆங்கில வழக்கின் அளவாய்வு (Survey of English Usage (SEU)) மற்றும் (2) 1975-இல் சுவீடனிலுள்ள லண்ட் பல்கலைக்கழகத்தில் ஜான் ஸ்வர்த்விக்கால் செயல்படுத்தபப்பட்ட பேச்சு ஆங்கிலத்தின் அளவாய்வு (Survery of Spoken Englsih (SSE)).

இதன் நோக்கம் சேகரிக்கப்பட்டு எழுத்துப்பெயர்ப்பு செய்யப்பட்ட பேச்சுச் சாதனங்களை இயந்திரம் படிக்கவியலும் வடிவில் வைத்திருக்கச் செய்வது ஆகும். உரை மாதிரிகள் கிட்டத்தட்ட 5 மில்லியன் பேச்சுச் சொற்களை உட்படுத்தும். பேச்சுச் சாதனம் இலக்கண ஆய்வின்றி, குறைக்கப்பட்ட எழுத்துப்பெயர்ப்பின் நிலை வழி சென்றுள்ளது. 1980-இன் தொடக்கத்தில் கணினிப்படுத்தப்பட்ட பேச்சு ஆங்கிலத்தின் லண்டன்-லண்ட் விரிதரவின் முதல் நகல் ஆய்விற்கும் புலன்விசாரணைக்கும் வேண்டி உலகம் முழுவதும் உள்ள ஆர்வமுள்ள ஆய்வாளர்களுக்குத் தரப்பட்டது.

2.2.2. பேச்சு ஆங்கிலத்தின் அளவாய்வு

பேச்சு ஆங்கிலத்தின் அளவாய்வு (Survey of Spoken English (SEU)) 200 மாதிரி உரைகளைக் கொண்டது. ஒவ்வொரு மாதிரியும் 5000 சொற்களைக் கொண்டிருக்கும்; மொத்தத்தில் 1 மில்லியன் சொற்களைக் கொண்டிருக்கும். பேச்சு மூல உரைகள் பின்வருவனவற்றை உள்ளடக்கும் ஒருவர் உரையாடல்கள் (உடனடியான, சொற்பொழிவு, விளையாட்டு மற்றும் விளையாட்டல்லா விமர்சனம், தயாரிக்கப்பட்ட ஆனால் எழுதப்படாத சொற்பொழிவு), (2) இருவர் உரையாடல் (நேருக்கு நேரான உரையாடல், இரகசியமாகப் பதிவு செய்யப்பட்டவை மற்றும் வெளிப்டையாகப் பதிவு செய்யப்பட்டவை, தொலைபேசி உரையாடல்) மற்றும் (3) (பேச்சுகள், நாடகங்கள், செய்திகள், எழுதப்பட்ட சொற்பொழிவுகள், கதைகள்).

விரிதரவானது பேச்சின் மீக்கூறு மற்றும் மொழிக்கு அப்பாற்பட்ட பண்புக்கூறுகளை அடையாளப்படுத்தும் ஒழுங்குமுறையால் எழுத்துப்பெயர்புச் செய்யப்பட்டுள்ளது. இச்செயற்பாட்டில் ஒவ்வொரு மீக்கூறும் மொழிக்கு அப்பாற்பட்ட பண்புக்கூறுகளும் அவ்வாறே கூர்ந்து கவனிக்கப்பட்டுக் குறிப்பிடப்பட்டுள்ளன. எல்லாப் பேச்சு உரைகளும் இலக்கண அடிப்படையில் ஆயப்பட்டு விரிதரவுக்குள் முறையாக உருப்படுத்தம் செய்யப்பட்டுள்ளன.

2.2.3. இயந்திரத்தால் படிக்கவியலும் பேச்சு ஆங்கில விரிதரவு

பேச்சு ஆங்கில விரிதரவு தற்காலப் பேச்சு பிரிட்டிஷ் ஆங்கிலத்தின் கிட்டத்தட்ட 100000 சொற்களைக் கொண்ட இயந்திரத்தால் படிக்கவியலும் விரிதரவாகும். இது 1984-85-இல் லங்காஸ்டர் பல்கலைக்கழகத்தின் மனிதவியல் ஆய்வு நிதி (Humanities Research Fund) மற்றும் ஐபிம் யுகே லிமிடெட் (IBM UK Ltd) என்பவைகளால் ஆதரவு தரப்பட்டு உருவாக்கப்பட்டது. இதன் நோக்கம் மனிதப் பேச்சில் இசையோட்டம் தருவதற்கான திட்டங்களை ஆய்வதற்கும் பரிசோதனை செய்வதற்கும் வேண்டி பயன்படுத்தவியலும் இயல்பான பேச்சு பிரிட்டிஷ் ஆங்கிலத்தின் மாதிரிகளைச் சேகரிப்பதாகும்.

பேச்சுக் கூட்டிணைப்புக்கு வேண்டி ஒரு மாதிரியாகப் பொருத்தமுறும் பேச்சு ஆங்கிலத்தின் வகையின் பெருமளவு மாதிரியைச் சேகரிக்க முக்கியத்துவம் தரப்பட்டது. எனவே உயர்ந்த நடையைச் சார்ந்த பேச்சின் சிறிய மாதிரிகள் (அதாவது செய்யுளைப் படித்தல் அல்லது சமயச் சொற்பொழிவு) உட்படுத்தப்பட்டன. பெரும்பாலான உரைகள் பிஎன்சி விரிதரவிலிருந்து எடுக்கப்பட்டுள்ளன. இது விமர்சனங்களிலிருந்து உரைகள், செய்தி ஒலிபரப்பு, பொது கேட்போர் கூட்டத்திற்கான சொற்பொழிவுகள், வரையறுக்கப்பட்ட கேட்போர் கூட்டத்திற்கான சொற்பொழிவுகள், சமய ஒலிபரப்பு, பத்திரிக்கை நடை அறிவிப்பு, கதை படித்தல், செய்யுள் ஒப்பித்தல், இருவர் உரையாடல், பிரச்சாரம் மற்றும் பிற பேச்சுக்கள் என்பனவற்றை உட்படுத்தும். விரிதரவு எழுத்துவடிவ அடிப்படையிலும் மீக்கூறு அடிப்படையிலும் எழுத்துப்பெயர்ப்பு செய்யப்பட்டுள்ளது: மீக்கூறின் மீது தொடரியலின் பாதிப்பின் ஆய்வை ஆனுமதிக்க வேண்டி CLAWS சொல் அடையாளப்படுத்தும் ஒழுங்குமுறையைப் பயன்படுத்தி இலக்கணத்திற்காக அடையாளப்படுத்தப்பட்ட வடிவம் உருவாக்கப்பட்டது. இவ்விரிதரவு பேச்சுக் கூட்டிணைப்பில் அல்லது பேச்சு அறிதல் களங்களில் ஆய்வுசெய்யும் மக்களுக்குப் பயனுள்ளது என்று நிருபிக்கப்பட்டுள்ளது. இது இயல்பான பேச்சு ஆங்கிலத்தின் ஒலியியலின் நெருங்கிய ஆய்வுக்கு மாணவர்களுக்குச் சந்தர்ப்பம் தந்து கற்பித்தல் நோக்கங்களுக்கு மதிப்புள்ள மூலவளமாக இருப்பது நிருபிக்கப்பட்டுள்ளது.

2.2.4. நியூசிலாந்து பேச்சு ஆய்கிலத்தின் வெலிங்டன் விரிதரவு

பேச்சு நியூசிலாந்து ஆங்கிலத்தின் வெலிங்டன் விரிதரவு (Wellington Corpus of Spoken New Zealand Eglish (WCSNZE)) 1994-இல் ஜானட் ஹோம்சின் வழிகாட்டின் கீழ் வில்லிங்டன்

பல்கலைக்கழகத்தில் உருவாக்கப்பட்டது. ஒவ்வொரு பகுதியும் ஒரு மில்லியன் சொற்களைக் கொண்டிருக்கவேண்டும் என்று ஒத்துக்கொள்ளப்பட்டது. நியூசிலாந்து பேச்சு விரிதரவு 1875-க்கும் 1992-க்கும் இடையில் சேகரிக்கப்பட்ட முறையான பேச்சு/ஒருவர்பேச்சு (12%), பகுதி முறையான பேச்சு / பெறப்பட்ட ஒருவர் பேச்சு (13%), முறைசாராப் பேச்சு / உரையாடல் (75%) என்பனவற்றை உள்ளடக்கும்.

சுருக்கங்கள் பேச்சின் ஒவ்வொரு வகையையும் காணப்படும் சூழல்களின் ஒரு பரப்பெல்லையை உட்படுத்தும் 15 வகுப்புகளாகப் பிரிக்கப்பட்டுள்ளன. இது ஒலிபரப்புச் செய்திகள், ஒலிபரப்பு ஒருவர்பேச்சுகள், வானிலை அறிக்கை ஒலிபரப்பு, விளையாட்டு விமர்சனம், நீதிபதிகளின் சுருக்க உரை, விரிவுரைகள், ஆசிரியர் ஒருவர்பேச்சுகள், நேருக்கு நேரான உரையாடல்கள், தொலைபேசி உரையாடல்கள், வாய்மொழி வரலாற்று நேர்காணல்கள், சமுதாய வழக்கு நேர்காணல்கள், வானொலிப் பேச்சுகள், நேர்காணல்களின் ஒலிபரப்பு, பாராளுமன்ற வாதங்கள், ஒப்பந்த நடவடிக்கைகைகள் மற்றும் கூட்டங்கள் மற்றும் பிறவற்றை உள்ளடக்கும்.

2.2.5. எடின்பர்க் பல்கலைக்கழக பேச்சு கால ஆணவமும் ஆங்கில மொழியின் விரிதரவும்

எடின்பர்க் பல்கலைக்கழக பேச்சு கால ஆணவமும் ஆங்கில மொழியின் விரிதரவும் (Edinburgh University Speech Timing Archive and Corpus of English (EUSTACE)) என்பது ஒலியியல் ஆய்வாளர்களுக்கும் பேச்சு கூட்டிணைப்பிலும் புரிந்துகொள்ளலிலும் ஈடுபட்டுள்ள பேச்சுத் தொழிநுட்பவியலார்களுக்கும் பயனுள்ள நல்ல மூலவளமாகும். இவ்விரிதரவு எடின்பரோ பல்கலைக்கழகத்தின் கோட்பாடு மற்றும் பயன்பாட்டு மொழியியல் துறையில் பதிவுசெய்யப்பட்ட 4608 பேச்சு வாக்கியங்களைக் கொண்டது. பிரிட்டிஷ் ஆங்கிலத்தின் 6 பேசுபவர்களால் பேசப்பட்ட இவ்வாக்கியங்கள் பேச்சில் காலத்தின் பல விளைவுகளைப் பரிசோதிப்பதற்கு வேண்டி வடிவமைக்கப்பட்டுள்ளது; இவை நீட்சிக்கும் ஒலிப் பொருளடக்கத்திற்கும் வேண்டி கட்டுப்படுத்தப்பட்டுள்ளன. ஒவ்வொரு வாக்கியத்திலும் முக்கியச் சொற்களின் துணைப் பகுதிகள் கண்டுபிடிக்கப்பட்டுள்ளன; அவை சில வாக்கியங்களின் மீக்கூறு உருப்படுத்தத்தைப் பற்றிய குறிப்புகளை உட்படுத்தும். எடுத்துக்காட்டு வாக்கியங்கள் இணைய தளத்தில் கேட்பதற்காக வைக்கப்பட்டுள்ளது.

2.2.6. கொரியன் பேச்சு விரிதரவு

கொரியன் தொலைபேசி உரையாடல் விரிதரவு (Korean Telephone Conversations speech Corpus (KTCSC)) தொடக்கத்தில் யுஎஸ்ஏ-இல் நண்பர்களை அழை ஆய்வுத்திட்டத்தின் (call-friend project) பகுதியாகப் பதிவு செய்யப்பட்டது. மொழி அடையாளம் காணல் ஆய்வுத் திட்டத்தின் (Language Identification (LID) Project)) ஆதரவில் LDC-ஆல் உரையாடல்கள் சேகரிக்கப்பட்டது. இத்தரவுமையம் கொரியன் தாய்மொழி பேசுபவர்களுக்கு இடையில் 100 தொலைபேசி உரையாடல்களைக் (மொத்தம் 25251 சொற்களைக் கொண்ட) கொண்டிருக்கின்றது. இதில் கிட்டத்தட்ட 44 மணி நேர செவிப்புலப் பேச்சு ஆவணங்கள் உள்ளது.

ஒவ்வொரு சொல்லும் ஐந்து தகவல் களங்களைக் கொண்டது: (1) ஹங்குலில் இருந்து எழுத்துவடிவம், (2) யேல் ரோமனய்சேசனில் இருந்து எழுத்து வடிவம், (3) உச்சரிப்பு, (4) கொரியன் தொலைப்பேசி உரையாடல் எழுத்துப்பெயர்ப்புகளில் சொற்களின் நிகழ்வெண், (5) சொற்களின் உருபனியல் பகுப்பாய்வு.

பகுதி 3

விரிதரவின் கருத்துருசார் பாகுபாடு

4.0. அறிமுகம்

மின் விரிதரவு புதிய கருத்துருவாகும். இதை எவ்வாறு வகைப்படுத்துவது என்பதில் பொதுவாக ஏற்றுக்கொள்ளத்தக்கக் கருத்து இல்லை. இருப்பினும் அவற்றை வகைப்படுத்த வேண்டுவது அவசியமாகும். விரிதரவுகளை வகைப்படுத்த பல காரணிகளைப் பயன்படுத்தலாம். மொழியியல் காரணிகள் அக அடிப்படையிலும் புற அடிப்படையிலும் அமையும். புறக் காரணிகள் பங்கெடுப்பாளர்கள், நேர்வு, சமூக நிலை, மொழியின் கருத்து பரிமாற்றச் செயல்பாடு போன்றவற்றுடன் தொடர்புடைய உரை வகைப்பாட்டியல் (text typology) அடிப்படையில் அமையும். அகக் காரணிகள் மொழியின் சிறுபகுதிகளுக்குள் மொழி அமைப்பொழுங்குகளின் மறுநிகழ்வு அடிப்படையில் அமையும். இவ்வெல்லாச் சிக்கல்களையும் எடுத்துக் கொண்டு விரிதரவைப் பின்வரும் வழியில் பரந்த முறையில் வகைப்படுத்தலாம்:

1. பனுவலின் இனம் (Genera of text)

2. தரவின் இயல்பு (Nature of Data)
3. பனுவலின் வகை (Type of Text)
4. திட்ட வரைவின் நோக்கம் (Purpose of design)
5. பயன்பாட்டின் இயல்பு (Nature of application)

3.1 பனுவலின் இனம் அடிப்படையில்

பனுவலின் இனம் அடிப்படையில் விரிதரவை எழுதப்பட்ட விரிதரவு (Written Corpus) என்றும், பேச்சு விரிதரவு (speech corpus) என்றும், பேச்சுசார் விரிதரவு (Spoken Corpus) என்றும் வகைப்படுத்தலாம்.

3.1.1 எழுத்து விரிதரவு

எழுத்து விரிதரவு எழுதப்பட்ட, அச்சடிக்ககப்பட்ட, வெளியிடப்பட்ட மற்றும் மின் அணு வடிவில் உள்ள பல மொழித் தரவுகளிலிருந்து சேகரிக்கப்பட்டவையாகும். எடுத்துக்காட்டாக, MIT Corpus of Indian Language என்பதைக் கூறலாம்.

3.1.2 பேச்சு விரிதரவு

இது முறையான மற்றும் முறைசாராத விவாதங்கள், வாதங்கள், முன்தயாரிக்கப்பட்ட பேச்சுகள், சாதாரணமான மற்றும் இயல்பான பேச்சுகள், இருவர் உரையாடல்கள், ஒருவர் உரையாடல்கள், மேடைப் பேச்சுகள் போன்றவைக் கொண்டிருக்கும். எடுத்துக்காட்டாக, Wellington Corpus of Spoken New Zealand English என்பது பேச்சு விரிதரவு ஆகும்.

3.2.3 பேச்சு சார் விரிதரவு

பேச்சுசார் விரிதரவு பேச்சுத் தரவின் நுட்பமான நீட்சியாகும். இதில் பேச்சானது ஒலி பெயர்ப்பிற்காக மட்டுமேயல்லாமல் பிற மாற்றங்கள் செய்யப்படாமல் எழுத்து வடிவில் உருப்படுத்தப்பட்டுள்ளன. எடுத்துக்காட்டாக, London-Lund Corpus of Spoken English என்பதைக் கூறலாம்.

3.2. தரவின் இயல்பு அடிப்படையில்

தரவின் இயல்பு அடிப்படையில் விரிதரவுகளைப் பொது விரிதரவு (General Corpus), சிறப்பு விரிதரவு (Special Corpus), துணைமொழி விரிதரவு (Sub Language Corpus), மாதிரி விரிதரவு (Sample Corpus), இலக்கிய வீரிதரவு (Literary Corpus), கட்டுப்பட்ட விரிதரவு (Monitor Corpus) என்று வகைப்படுத்தலாம்.

3.2.1. பொது விரிதரவு

பொது விரிதரவு வேறுபட்ட துறைகள், இனங்கள், பாடக் களங்கள் மற்றும் நடைகள் இவற்றைச் சார்ந்த பொது உரைகளைக் கொண்டிருக்கும். எடுத்துக்காட்டாக British National Corpus என்பதைக் கூறலாம். இதன் வடிவம் மற்றும் பயன்பாடு அடிப்படையில் இது உரை சேகரிப்பில் முடிவான எண்ணிக்கையைக் கொண்டது. அதாவது, இதில் உரை வகைகளின் மற்றும் சொற்கள் மற்றும் வாக்கியங்கள் இவற்றின் எண்ணிக்கை ஒரு எல்லைக்குட்பட்டது. இது கால அடிப்படையில் வளரவோ புதிய உரைகளின் இருப்பு அடிப்படையில் மேலும் அதிகரிக்கப்படவோ செய்யாது. ஆனால் இது அளவில் மிகப் பெரிதாகவும் வகைவேறுபாட்டில் செழிப்பானதாகவும் உருப்படுத்தத்தில் விரிந்ததாகவும் பயன்பாட்டில் பரந்தாகவும் இருக்கவேண்டும்.

3.2 சிறப்பு விரிதரவு

சிறப்பு விரிதரவு மொழியின் குறிப்பிட்ட வகை, கிளைமொழி மற்றும் ஆய்வுத் தலைப்பின் சில பண்புகளை வலியுறுத்தும் பாடம் இவற்றிற்கு வேண்டி பொது விரிதரவில் மாதிரிபடுத்தப்பட்ட உரைகளிலிருந்து திட்டவரைவுசெய்யப்பட்டுள்ளது. எடுத்துக்காட்டாக CHILDES Data Base என்பதைக் கூறலாம். இது நோக்கம் அடிப்படையில் அதன் அளவிலும் உள்ளடக்கத்திலும் வேறுபடும். இது அசாதாரணமான பண்புக்கூறுகளின் கூடுதல் பகுதியைக் கொண்டிருப்பதால் இது மொழியின்

வருணனைக்குப் பங்களிப்பு செய்யாது. இது இயல்பான நடத்தையில்லாத மக்களிடமிருந்து தரவுகளைப் பதிவு செய்வதால் இதன் மூலம் நம்பத்தகுந்ததல்ல. சிறப்பு விரிதரவு (அதன் தரப்பட்ட நோக்கத்திற்குள் அல்லாமல்) சரிசமமானதல்ல; பிற நோக்கத்திற்குப் பயன்படுத்தினால், மொழிப்பகுதிகளின் பிழையான மற்றும் மாற்றப்பட்ட பார்வையைத் தரும். இது இயல்பான, நம்பகமான மொழியின் ஒன்றோ மற்றொன்றோ வகையின் பண்புக்கூறுகளை வெளிப்படுத்துவதால் இது கொள்கை அடிப்படையில் வேறுபட்டதாகும். மொழியின் பிரதிநித்துவமல்லாத இயல்பை உட்படுத்துவதன் காரணமாக குழந்தைகள், தாய்மொழி அல்லாதவர், புறக் கிளைமொழிகளைப் பயன்படுத்துவோர் என்போரின் மொழி மற்றும் கருத்துப்பரிமாற்றத்தின் மிகச் சிற்பான பரப்புகள் (எ.கா. ஏலம், மருத்துவப் பேச்சு, சூதாட்டம், நீதிமன்ற நிகழ்வு) என்பனவற்றின் விரிரவு சிறப்பு விரிதரவுகளாகக் கருதப்படவேண்டும்.

3.5.7. துணை மொழி விரிதரவு

துணை மொழி விரிதரவு ஒரு குறிப்பிட்ட மொழியின் ஒரே ஒரு உரை வகையை கொண்டிருக்கும். இது ஒரு குறிப்புரை விரிதரவிலிருந்து (reference corpus) மொழிக் கற்றையின் மறு நுனியில் இருக்கின்றது. இதன் ஒருபடித்தான அமைப்பும் சிறப்பிக்கப்பட்ட சொற்களஞ்சியமும் மாதிரி அடிப்படையில் நல்ல மற்றும் நெருங்கிய தன்மைகளை நிறுவ வேண்டி தரவின் அளவை சிறதாக வைக்க அனுமதிக்கின்றது.

3.5.8. மாதிரி விரிதரவு

இது சிறப்புத் தன்மையான விரிதரவின் ஒரு வகையாகும். இது மிகுந்த கவனத்தோடும் விரிவாகவும் ஆயப்பட்ட உரைகளின் முற்றுநிலைச் சேகரிப்பைக் கொண்டிருக்கும் சிறிய மாதிரிகளால் உருவாக்கப்பட்டதாகும். எடுத்துக்காட்டாக Zurich Corpus of English News Papers என்பதைக் கூறலாம். மாதிரி விரிதரவு ஒருதடவை உருவாக்கப்பட்டுவிட்டால் இதில் ஏதாவது சேர்க்கப்படவோ எந்த வழியிலாவது மாற்றப்படவோ செய்வதில்லை; ஏனென்றால் எந்த மாற்றமும் இதன் உள்ளடக்கத்தை (தரவுமையத்தை) சரிசமமில்லாமல் செய்வதுடன் அதன் ஆய்வுத் தேவையை சிதைக்கும். உரைகளைப் பொறுத்தவரையில் மாதிரிகள் குறைந்த எண்ணிக்கையிலானவை மற்றும் நிரந்தர வடிவ அளவு உள்ளது. எனவே அவை உரைகளாகத் தகுதிபெறாது.

3.5.9. இலக்கிய விரிதரவு

இது மாதிரி விரிதரவின் ஒரு சிறப்பான வகையாகும். இதில் பல வகைகள் இருக்கின்றன. பைபிள்சார் மற்றும் இலக்கிய ஆய்வுகள் வெகுநாட்களுக்கு முன்பே விரிதரவு மொழியியல் என்ற துறையை தொடங்கியது. ஆக்கியோரின் படைப்புகளின் ஒழுங்கை நிறுவுதல் போன்ற இலக்கியக் களத் திறன்கள் இருக்கின்றன. இம்மாதிரியான விரிதரவு உருவாக்கத்தின் வகைப்படுத்தும் அளவீடுகள் ஆக்கியவர், இனம் (எ.கா. தூது சிறுகதைகள், கதைகள், போன்றவை), குழு (எ.கா. காதல் புலவர்கள், அகஸ்டன் உரைநடை எழுத்தாளர்கள், விக்டோரியன் கதை ஆசிரியர்கள், போன்றோர்), பொருள் (எ.கா. புரட்சி எழுத்துகள், குடுப்பக் கூற்றுகள், தொழில்மயமாக்கம் போன்றவை) மற்றும் மதிக்கத்தக்க காரணிகளாகப் பிற சிக்கல்கள் என்பனவற்றை உள்ளடக்கும். இருப்பினும் அறியப்பாத காரணத்திற்காக நாடகங்களிலிருந்து உருவாக்கப்பட்ட விரிதரவு உரைநடையிலிருந்தும் செய்யுளிலிருந்தும் தனியாக வைக்கப்படுகின்றது.

3.5.10. கட்டுப்படுத்தும் விரிதரவு

கட்டுப்படுத்தும் விரிதரவு (Monitor Corpus) மொழியின் மாற்றத்தைப் பிரதிபலிக்கும் தரவின் நிரந்தர சேர்க்கைக்கு வழி செய்யும் உரைகளின் வளர்கின்ற எல்லையற்ற தொகுப்பாகும். விரிதரவின் நிரந்தர வளர்ச்சி மொழியில் மாற்றத்தை பிரதிபலிக்கும். ஒரே இணைப்பாக்கத் திட்டவரைவு ஆண்டுக்கு ஆண்டு பின்பற்றப்படுகின்றது. கட்டுப்படுத்தும் விரிதரவின் அடிப்படை ஒரு தனி வருடத்தில் பேசப்பட்ட அல்லது எழுதப்பட்ட உரைகளுக்குக் குறிப்புரை செய்வதாகும். கட்டுப்படுத்தும் விரிதரவிலிருந்து நாம் புதிய சொற்களைக் கண்டுபிடிக்கலாம்; வழக்கிலுள்ள வேறுபாட்டைத் தொடர்ந்து அறியலாம்; பொருண்மையிலுள்ள வேறுபாட்டை உற்றுநோக்கி அறியலாம்; நிகழ்வெண்ணின் நீண்ட கால நியமத்தை நிலைநிறுத்தலாம்; சொல் தகவலின் விரிந்த எல்லைப்பரப்பை ஆக்கலாம். தரவின் புதிய மூலங்கள் வருவதாலும் புதிய செயல்முறைகள் அரிதான சாதனங்களை செழிப்பாக இருக்கும் வண்ணம் செய்வதாலும் நாளடைவில் கட்டுப்படுத்தும் விரிதரவின் சரிசமநிலை மாறும். இவ்வொழுக்கின் விகிதம் அவ்வவ்போது நேர் செய்யப்படும்.

3.6. உரையின் வகை அடிப்படையில்

உரையின் வகை அடிப்படையில் விரிதரவை ஒருமொழிய விரிதரவு, இரு மொழிய விரிதரவு, பன்மொழிய விரிதரவு என்று வகைப்படுத்தலாம்.

3.3.1. ஒருமொழிய விரிதரவு

ஒருமொழிய விரிதரவு (எ.கா. ISI Bangia Corpus) ஒரு காலத்திலோ பலகாலத்திலோ பயன்பாட்டை பிரதிநுத்துவம் செய்யும் ஒரு மொழியின் பிரதிநித்துவ உரைகளைக் கொண்டிருக்கும்.

3.3.2. இருமொழிய விரிதரவு

இரு உறவுள்ள அல்லது உறவற்ற மொழிகளின் விரிதரவுகள் ஒரே சட்டகத்தில் வைக்கப்படும் போது இருமொழிய விரிதரவு (எ.கா. ISI Bangla Corpus) உருவாகும். இம்மொழிகள் இன அடிப்படையில் அல்லது வகைப்பாட்டியல் அடிப்படையில் உறவுள்ளவை என்றால் அவை இணை விரிதவுகளாக மாறும்; உரைகள் சில முன்னரே வரையறுக்கப்பட்ட காரணிகளைப் பின்பற்றி வரிசைப்படுத்தப்படும். இங்கு வடிவ அளவு, பொருளடக்கம் மற்றும் களம் என்பன விரிதரவுக்கு விரிதரவு மாறும். இணை விரிதரவின் நேர்வில் இது அனுமதிக்கப்படுவதில்லை.

3.3.3. பன்மொழிய விரிதரவு

பன்மொழிய விரிதரவு (எ.கா. Crater Corpus) இரண்டிற்கும் மேற்பட்ட மொழிகளிலிருந்து நல்ல பிரதிநித்துவச் சேகரிப்புகளைக் கொண்டிருக்கும். பொதுவாக இங்கும் இருமொழிய விரிதரவிலும் உரைகள் வேறுபட்ட மொழிகளைச் சார்ந்திருந்தாலும் ஒற்றுமையுள்ள உரை வகைப்பாடுகள் மற்றும் ஒரே மாதிரியாக்க நெறிமுறைகள் பின்பற்றப்படுகின்றன.

3.7. திட்டவரைவின் நோக்கம் அடிப்படையில்

திட்டவரைவின் நோக்கம் அடிப்படையில் விரிதரவை குறிப்புகள் தரப்படாத (unannotated corpus), குறிப்புகள் தரப்பட்ட விரிதரவு (annotated corpus) என்று வகைப்படுத்தலாம்.

3.4.1. குறிப்புகள்தரப்படாத விரிதரவு

குறிப்புகள் தரப்படாத விரிதரவு (e.g. MIT Corpus of Indian Languges) மொழிசார் அல்லது மொழிசாராத் தகவல்களைக் கொண்டிராமல் உரைகளின் எளிய தொடக்க நிலையை பிரதிநித்துவம் செய்யும். இது மொழி ஆய்விற்கு ஓரளவுக்கு பயனுள்ளதாகக் கருதப்படுகின்றது; இருப்பினும் விரிதரவின் பயன்பாடு குறிப்புகள் தரப்படும் போது போதிய அளவு அதிகரிக்கும்.

3.4.2. குறிப்புகள் தரப்பட்ட விரிதரவு

குறிப்புகள் தரப்பட்ட விரிதரவு (எ.கா. British National Corpus) சில கூடுதலான தகவலகளை (ஆய்வுக் குறிகள், சொல்வகைப்பாட்டுக் குறிகள், இலக்கண வகைப்பாட்டுத் தகவல்கள்) உரையில் பதிவுசெய்ய வேண்டி திட்டவரைவு செய்பவர்களால் வெளியிலிருந்து செருகப்பட்ட அடையாளங்களையும் குறிமங்களையும் கொண்டிருக்கும். குறிப்புகள் தரப்படாத விரிதரவுக்கு மாறாக குறிப்புகள் தரப்பட்ட விரிதரவு உருபனியல் பகுப்பாய்வு, வாக்கியப் பகுப்பாய்வு, தகவல் மீட்பு, சொற்பொருண்மை மயக்கம் நீக்கல், இயந்திர மொழிபெயர்ப்பு போன்றவற்றை உள்ளடக்கிய மொழித் தொழில் நுட்பத்திற்கு வேண்டிய பல செயல்பாடுகளில் பயன்படும் பொருத்தமான தகவல்களைத் தருவதற்கு கூடுதல் தகுதியானது.

3.8. பயன்பாட்டின் இயல்பு அடிப்படையில்

பயன்பாட்டின் இயல்பு அடிப்படையில் விரிதரவை வரிசைப்படுத்தப்பட்ட விரிதரவு (Aligned Corpus), இணையான விரிதரவு (Parallel Corpus), நோக்கீட்டு விரிதரவு (Reference Corpus), ஒப்பீட்டு விரிதரவு (Comparable Corpus), சந்தர்ப்பவாத விரிதரவு (Opportunistic Corpus) என்று வகைப்படுத்தலாம்.

3.8.2. வரிசைப்படுத்தப்பட்ட விரிதரவு

இது இரு அல்லது பன்மொழி விரிதரவின் ஒரு வகையாகும். இதில் ஒரு மொழியின் மற்றும் அதன் மொழி பெயர்ப்புகளில் உரைகள் வாக்கியங்களுக்கு வாக்கியமாகவும் தொடர்களுக்கு தொடராகவும் சொல்லுக்கு சொல்லாகவும் வரிசைப்படுத்தப் பட்டுள்ளன. எடுத்துக்காட்டாக The Canadian Hansard Corpus என்பதைக் கூறலாம்.

3.8.3. இணையான விரிதரவு

இணையான விரிதரவு அல்லது மொழி பெயர்ப்பு விரிதரவு இரண்டு முறை மொழி பெயர்ப்பு நிகரன்களின் பரிசோதனையை அனுமதிக்கும் மொழிகளின் உரைகளையும் மொழி பெயர்ப்புகளையும் கொண்டிருக்கிறது. இதில் ஒரு மொழியில் உள்ள உரைகள் மற்றும் அவற்றின் மொழி பெயர்ப்புகள் வாக்கியங்களுக்கு வாக்கியம், தொடர்களுக்கு தொடர், சொல்லுக்கு சொல் வரிசைப்படுத்தப் பட்டுள்ளன. எடுத்துக்காட்டாக Chemnitz German – English / English – German Translation Corpus என்பதைக் கூறலாம்.

3.8.4. நோக்கீட்டு விரிதரவு

இது ஒரு மொழியின் மொழியை பற்றிய விரிவான தரவை தரும்படி திட்டமிடப்பட்டுள்ளது. இது ஒரு மொழியின் இலக்கணங்கள், அகராதிகள், சொற்கஞ்சியங்கள் மற்றும் பிற நோக்கீட்டு பொருள்கள் என்பனவற்றை உருவாக்கும் படி தேவையான மொழி வகைகளையும் சிறப்பான சொற்றொகையையும் பிரதிநுத்துவம் செய்வதைக் குறிக்கோளாக கொண்டது. எடுத்துக்காட்டாக Bank of English என்பதைக் கூறலாம்.

3.8.5. ஒப்பீட்டு விரிதரவு

இது ஒன்றுக்கு மேற்பட்ட மொழிகளின் அல்லது மொழி வகைகளின் ஒற்றுமையுள்ள உரைகளின் சேகரிப்பு ஆகும். இது ஒரு வகையான பன்மொழி பெருந்தரவாகும். இதில் பொருளடக்கம், இனம் மற்றும் நடை இவற்றில் வேறுபட்ட மொழிகளின் உரைகளை கொண்டுள்ளது. எடுத்துக்காட்டாக Corpus of European Union என்று கூறலாம்.

3.8.6. சந்தர்ப்பவாத விரிதரவு

சந்தர்ப்பவாத விரிதரவு பல முறைகளிலும் கிடைக்கும் மின் உரைகளின் சேகரிப்பாகும். இது பெரும்பாலும் முடிவுராமல் இருக்கும். இது பெரும்பாலும் முற்றுப்பெறாமலும் முழுமை பெறாமலும் இருக்கும். கட்டுப்படுத்தும் விரிதரவு பெரும்பாலும் சந்தர்ப்பவாத விரிதரவாகக் கருதப்படும்.

பகுதி 4

எழுதப்பட்ட விரிதரவின் உருவாக்கம்

4.0. அறிமுகம்

விரிதரவை திட்டவரைவு செய்வதிலும் உருவாக்குவதிலும் நிர்வகிப்பதிலும் பல சிக்கல்கள் இருக்கின்றன. விரிதரவை உருவாக்குவதிலும் செயற்பாங்குசெய்வதிலும் உள்ள சிக்கல்கள் விரிதரவின் வகை மற்றும் பயன்பாட்டின் நோக்கம் இவற்றின் அடிப்படையில் வேறுபடும். பேச்சு விரிதரவு உருவாக்குவதுடன் தொடர்புடைய சிக்கல்கள் உரை விரிதரவு உருவாக்குவதுடன் தொடர்புடைய சிக்கல்களிலிருந்து வேறுபடும். எனவே விரிதரவு உருவாக்கத்தை இரண்டாகப் பகுக்கலாம்:

1. எழுத்து உரை விரிதரவை உருவாக்குதல்
2. பேச்சு விரிதரவை உருவாக்குதல்

பேச்சு விரிதரவுகளை உருவாக்க பின்வரும் கருத்துக்கள் கவனத்தில் கொள்ளப்பட வேண்டும்:

1. பயன்பாட்டின் நோக்கம்
2. தகவலாளியின் தேர்வு
3. விரிதரவின் அளவு
4. அமைப்புகளின் தேர்வு
5. தரவு மாதிரியின் முறை
6. தரவு சேகரித்தலின் முறை
7. எழுத்துப்பெயர்ப்பின் சிக்கல்
8. தரவு குறியாக்கத்தின் வகை
9. தரவு கோப்புகளின் நிர்வகிப்பு
10. உள்ளீட்டு தரவின் திருத்தல்

11. உரைகளின் பகுப்பாய்வு
12. உரைகளின் ஆய்வு

எழுத்து உரை விரிதரவை உருவாக்குவது பின்வரும் விஷயங்களை உள்ளடக்கும்:

1. விரிதரவின் அளவு
2. மூல மாதிரி
3. தாய்மொழித் தன்மை
4. இலக்கு பயன்படுத்துபவரின் நிர்ணயம்
5. காலகட்டத்தின் தேர்வு
6. ஆவணங்களின் தேர்வு
7. உரை ஆவணங்களின் சேகரிப்பு (நூல்கள், செய்திதாள்கள், காலஇதழ்கள் போன்றவை)
8. தரவு மாதிரியின் நெறிமுறை (ஒருவருடைய தேவைக்கு தக்கவாறு சேகரிக்கப்பட்ட விஷயங்களின் வரிசை)
9. தரவு உள்ளீட்டின் முறை (ஒழுங்கற்ற, சீரான, தேர்வு முறை)
10. விரிதரவைத் திருத்துதல் (தவறுகளைத் திருத்தல், அயற்சொற்கள், மேற்கோள்கள், கிளைமொழிகள் இவற்றை விட்டுவிடுதல்)
11. விரிதரவு கோப்புகளின் நிர்வகிப்பு
12. வெளியீட்டு உரிமையின் சிக்கல்

4.1. விரிதரவின் வடிவ அளவு

ஒரு விரிதரவு எவ்வளவு பெரியதாக இருக்க வேண்டும் என்பது கேள்விக்குறியாகும். விரிதரவு உருவாக்கத்தில் வடிவ அளவு முக்கிய பங்குவகிக்கிறது. இது விரிதரவில் மொத்தம் எத்தனை சொற்கள் (tokens) இருக்கின்றன, வேறுபட்ட சொற்கள் (types) இருக்கின்றன என்பதுடன் தொடர்புடையதாகும். இது விரிதரவில் எத்தனை வகைபாடுகள் வைக்கப்பட வேண்டும், ஒவ்வொரு வகைப்பாடடிலும் எத்தனை உரைகள் மாதிரிகள் வைக்கப்பட வேண்டும், ஒவ்வொரு மாதிரியிலும் எத்தனை சொற்கள் இருக்க வேண்டும் என்ற தீர்மானத்தை உள்ளடக்கும்.

ஒரு விரிதரவில் எத்தனைச் சொற்கள் அல்லது சொல் வகைகள் இருக்க வேண்டும் என்பது முக்கியமாக இருந்தாலும் தற்போது மிகப் பெரிய விரிதரவு தான் மொழியைப் பிரதிபலிப்பதாக அமைகின்றது என்பது நம்பப்படுகிறது. எடுத்துக்காட்டாக, Bank of Enlish, BNC, Cobuild Corpus, Longman/ Lancaster Corpus போன்ற விரிதரவுகள் 100 மில்லியன் சொற்களைக் கொண்டிருக்கின்றன.

4.2. உரையின் பிரதிநிதித்துவத் தன்மை அல்லது மூலமாதிரி

போகப்போக உரைகளின் பிரதிநிதித்துவத் தன்மையின் சூழலில் விரிதரவின் வடிவ அளவு முக்கியத்துவம் இழக்கும். மிகப்பெரிய தரவு ஒரு மொழியின் அல்லது அதன் ஒருவகையை ஒரு சிறிய ஆனால் நன்றாகச் சமநிலையை உடைய விரிதரவைக் காட்டிலும் கூடுதல் பிரதிநிதித்துவம் செய்யத் தேவையில்லை. உரைகளின் பெரிய சேகரிப்பு நாம் எந்த பொதுமையாக்கத்திலும் பயன்படுத்தவியலும் ஒரு விரிதரவாக இருக்கத் தேவையிலை. விரிதரவின் ஆய்வின் அடிப்படையிலான கண்டுபிடிப்புகள் ஒரு மொழியின் முழுமைக்கோ அதன் ஒரு குறிப்பிட்ட பகுதிக்கோ பொதுமையாக்கம் செய்யப்பட்டால் தான் நாம் ஒரு விரிதரவை பிரதிநிதித்துவத் தன்மையுடையது என்று கூறவியலும். எனவே தான் நாம் விரிதரவில் தரவின் அளவின்மீது கவனக்குவிப்பு செய்யாமல் தரவின் பண்பிற்கு முக்கியத்துவம் தருகின்றோம். அதாவது தரவு ஒரு விரிதரவுக்குள் மொழிப்பயன்பாட்டின் எல்லா சாத்தியமான பரப்புகளிலிருந்தும் சரியான விகிதத்தில் இருக்கவேண்டும். விரிதரவின் ஒட்டுமொத்த வடிவ அளவு பொருத்தமான பிரதிநித்துவத்தன்மையைப் பெற உதவ வேண்டி மூலப்பொருள்களின் வேறுபட்டத்தன்மையைப் பிரதிபலிக்க வேண்டும். உரை வகைபாடுகளுக்குள் எந்த அளவுக்கு தனிநிலை மாதிரிகளின் எண்ணிக்கை அதிகமாக இருக்கிறதோ அந்த அளவுக்கு மொழியியல் மாறிகளின் ஆய்வின் நம்பகத்தன்மை அதிகமாக இருக்கும். எடுத்துக்காட்டாக, Brown Corpus, LOB Corpus, SEU என்பன அமேரிக்காவிலும் யுகேயிலும் பயன்படுத்தப்படும் ஆங்கில மொழியின் நல்ல பிரதிநிதிகளாக நாம் கருதவியலும் அளவுக்கு கவனமாகத் திட்டவரைவு செய்யப்பட்டுள்ளன. இருப்பினும் மிகக்கூடுதல் வேறுபட்ட அமைப்பையும் பிரதிநிதித்துவச் சட்டகத்தையும் கொண்ட 100 மில்லியன் சொற்களையுடைய BNC விரிதரவுடன் Brown, LOB மற்றும் SEU விரிதரவுகளை ஒப்பிடும்போது எந்த அளவிற்கு இவ்விரிதரவுகள் பொருளடக்கத்தில் சிறிதாகவும் அமைப்பில் குறைந்த வேறுபாடுடையதாகவும் இருக்கின்றன என்பதைக் காட்டும். இது விரிதரவில் வடிவ அளவு மற்றும் பிரதிநித்துவம் இவற்றிற்கிடையிலான சிக்கல்களை மிக எளிதாக அனுபவவாத அடிப்படையில் தீர்க்கின்றது.

4.3. தாய்மொழித் தன்மை

தாய்மொழிக் குறித்த கேள்வி விரிதரவு உருவாக்கும் போது தாய்மொழி பேசுபவரின் உரைகளைத் தேர்வு செய்ய வேண்டுமா அல்லது தாய்மொழி பேசாதவரின் உரைகனளத் தேர்வு செய்ய வேண்டுமா என்பது கேள்வி. பொதுவாக கூறப் போனால் தாய்மொழி பேசுபவருடைய உரைகள் தாய்மொழி பேசாதவர்களின் உரைகாட்டிலும் நம்பத் தகுந்ததாய் அமையும். இருப்பினும் அது ஒரு monitor corpus என்றால் தாய்மொழியாகப் பயன்படுத்தாதவர்களைக் காட்டிலும் தாய்மொழியாகப் பயன்படுத்துவோர்களால் உருவாக்கப்படும் உரைகள் முக்கியத்துவம் பெறவேண்டும். ஏனென்றால் மானிட்டர் விரிதரவின் நோக்கம் எல்லா மொழியியல் மற்றும் மொழித் தொழில் நுட்பச் செயல்பாடுகளிலும் முன்மாதிரியாகக் கருதப்படும் மொழியை பிரதிநித்துவம் செய்வதாகும். எடுத்துக்காட்டாக, பிரிட்டிஷ் ஆங்கில மொழியின் விரிதரவு உருவாக்கத்தின் போது இந்திய ஆங்கில உரைகள் அதில் உள்ளடக்க கூடாது. இது மறுதலையாகவும் அமையும்.

4.4. இலக்குப் பயன்பாட்டாளரின் நிர்ணயம்

ஒரு பொதுவான விரிதரவுக்கு குறிப்பட்ட இலக்கு பயன்பாட்டாளர் இல்லை. யாரும் ஒரு பொது விரிதரவை எதற்கும் பயன்படுத்தலாம். ஒரு சிறப்பு விரிதரவுக்கு (specialized corpus) யார் பயன்பாட்டாளர் என்ற கேள்வி முக்கியமானதாகும். ஒவ்வொரு விசாரணையாளருக்கும் ஆய்வாளருக்கும் குறிப்பிட்ட தேவை இருப்பதன் காரணமாக விரிதரவு அதனடிப்படையில் திட்டவரைவு செய்யப்படவேண்டும். எடுத்துக்காட்டாக மொழிபெயர்ப்பு கருவிகளின் உருவாக்கத்தில் ஈடுபட்டுள்ள ஒரு நபர் பொது விரிதரவைக் காட்டிலும் இணையான விரிதரவை வேண்டுவார். அதுபோல் இரண்டோ அதற்கு மேற்பட்ட மொழிகளின் ஒப்பீட்டு ஆய்வில் ஈடுபட்டுள்ள நபர் மானிட்டர் விரிதரவைக் காட்டிலும் ஒப்பிடவியலும் விரிதரவை வேண்டுவார்.

இலக்கு பயன்பாட்டாளர்	**விரிதரவு**
வர்ணனை மொழியியலார்கள்	பொதுவான விரிதரவு, எழுதப்பட்ட விரிதரவு மற்றும் பேச்சு பெருந்தரவு
பேச்சு தொழில் நுட்ப ஆய்வாளர்	பேச்சுப் விரிதரவு
அகராதியியலார்கள்' மற்றும் கலைச்சொல் அகராதியியலார்கள்	பொது விரிதரவு, மானிட்டர் விரிதரவு, சிறப்பு விரிதரவு, குறிப்பு விரிதரவு, சந்தர்ப்பவாத விரிதரவு போன்றவை
உரையாடல் ஆய்வாளர்கள்	பேச்சு விரிதரவு, அடையாளப்படுத்தப்பட்ட

	விரிதரவு, சிறப்பிக்கப்பட்ட விரிதவு
சமுதாய மொழியிலார்கள்	பொது விரிதரவு, பேச்சு விரிதரவு, எழுத்து விரிதரவு மற்றும் சிறப்பு விரிதரவு
உளமொழியியலார்கள்	சிறப்பிக்கப்பட்ட விரிதரவு, பேச்சு விரிதரவு, எழுத்து விரிதரவு
வரலாற்றியலார்கள்	இலக்கிய விரிதரவு, இருகால விரிதரவு
சமூக அறிவியலார்கள்	பொது விரிதரவு, பேச்சு விரிதரவு, எழுத்து விரிதரவு, சிறப்பு விரிதரவு
ஒப்பீட்டு மொழியியலார்கள்	இருமொழிய விரிதரவு, பன்மொழிய விரிதரவு, இணை விரிதரவு, ஒப்பிடவியலும் விரிதரவு, அடையாளப்படுத்தப்பட்ட விரிதரவு
மொழிபெயர்ப்பு வல்லுநர்கள்	இருமொழிய விரிதரவு, பன்மொழிய விரிதரவு, இணை விரிதரவு, ஒப்பிடவியலும் விரிதரவு, அடையாளப்படுத்தப்பட் விரிதரவு
தகவல் மீட்பு வல்லுனர்கள்	பொது விரிதரவு, மானிட்டர் விரிதரவு மற்றும் அடையாளப்படுத்தப்பட்ட விரிதரவு
அடையாளப்படுத்தும் பகுப்பாயும் பகுத்துக்குறிக்கும் வல்லுநர்கள்	அடையாளப்படுத்தப்பட்ட மானிட்டர், எழுதப்பட்ட விரிதரவு, பேச்சு விரிதரவு, பொது விரிதரவு
மையஇலக்கண திட்டவரைவாளர்கள்	ஒப்பிடவியலும் விரிதரவு, இருமொழிய விரிதரவு மற்றும் பொது விரிதரவு
சொல் பொருண்மையாக்கம் நீக்குபவர்கள்	அடையாளப்படுத்தப்பட்ட மானிட்டர் எழுதப்பட்ட பேச்சு பொது விரிதரவு
ஆசிரியர்கள், மாணவர்கள்	கற்பவர், மானிட்டர் பொது விரிதரவு

4.5. காலகட்டத்தின் தேர்வு

மொழி கால அடிப்படையில் மாறும். எனவே ஒரு குறிப்பிட்ட காலகட்டத்திற்கு மொழியின் பண்புக் கூறுகளைக் கண்டுபிடிக்க நாம் ஒரு குறிப்பிட்ட காலகட்டத்தை நிர்ணியக்க வேண்டும். பெருந்தரவு தெளிவான கால அடையாளப்படுத்தலுடன் ஒரு குறிப்பிட்ட காலத்தை உட்படுத்த முயற்சிக்கிறது. எடுத்துக்காட்டாக, இந்திய மொழிக்கான MIT விரிதரவு தற்கால மொழியின் நிலையைப் பிரதிநித்துவம் செய்யவேண்டி 1981-லிருந்து 1995 வரையிலான காலக்கட்டத்தில் வெளியிடப்பட்ட விஷயங்களைக் கொண்டிருக்கின்றது.

4.6. உரைகளின் வகையின் தெரிவு

எழுதப்பட்ட விரிதரவு திட்டமிடுதலின் முக்கியமான கேள்வி அது எழுத்து உரைகளின் எல்லா வகைகளையும் கொண்டிருக்க வேண்டுமா என்பதாகும். பெரும்பாலான விரிதரவுகள் தரப்படுத்தப்பட்ட எழுத்தாக்கங்களின் உரைகளைக் கொண்டிருப்பதை விரும்புகின்றது. ஒரு பொது விரிதரவின் நோக்கம் என்னவென்றால் எவை ஒரு மொழியின் மைய (பொதுவான) பண்புக்கூறுகள் எவை தனிப்பட்ட (சிறப்புப்) பண்புக்கூறுகள் என்பதைக் கண்டுபிடிப்பதாகும். எனவே ஒரு விரிதரவு தற்காலத்திய எழுத்தாக்கங்களில் மிக நல்லதை மட்டும் உள்ளடக்க வேண்டும் என்பதல்ல. அளவீடு அடிப்படையிலான விகிதம் போதுமானதாகும். இவ்வாறு ஒரு விரிதரவு மனித அறிவின் வேறுபட்ட கிளைகளை எடுக்கப்படும் விஷயங்களின் சேகரிப்பாகும். இங்கு மிகப் புகழ்ப்பெற்ற எழுத்தாளர்களின் எழுத்தாக்கங்களும் அறிமுகம் இல்லாத எழுத்தாளர்களின் எழுத்தாக்கங்களும் ஒன்றாகக் கருதப்படும். தரவுச் சேகரிப்பிற்காக வெளியீடுகளின் ஆணவங்களின் (நூல்கள், செய்தித்தாள்கள், காலவிழ்கள் போன்றவை) சேகரிப்புக்காக எல்லா கேட்டலாக்குகளும் பட்டியல்களும் பார்க்கப்படவேண்டும். இது குறிப்பிட்ட மூலத்தின் தனித்தன்மை மறைக்கப்படுகின்ற பல்வேறுபட்ட மூலங்களிலிருந்தும் துறைகளிலிருந்தும் விஷயங்களைச் சேகரிப்பதால் அதன் இயல்பில் பெரும்பான்மையும் பலபடித்தானதாக இருக்கின்றது. திரிக்கப்பட்ட பிரதிநிதித்துவத்தின் எந்த வகைக்கும் எதிராக விரிதரவில் வேறுபட்டத்தன்மை பாதுகாக்கப்படுகின்றது.

MIT–தமிழ் விரிதரவு இலக்கியத்திலிருந்து 20%யையும், நுண்கலையிலிருந்து 5%யையும், சமூக அறிவியலிருந்து 15%யையும் இயற்கை அறிவியலிருந்து 15%யையும் வணிகவியலிருந்து 10%யையும் மக்கள் தகவல் தொடர்விலிருந்துரு 30%யையும் மொழி பெயர்ப்பிலிருந்து 5%யையும் கொண்டிருக்கிறது. ஒவ்வொரு வகைப்பாடும் சில துணைவகைப்பாடுகளைக் கொண்டிருக்கிறது. எடுத்துக்காட்டாக,

- இலக்கியம் புதினங்கள், சிறுகதைகள், கட்டுரைகள் போன்றவற்றைக் கொண்டிருக்கிறது.
- நுண்கலைகள் ஓவியங்கள், வரைபடங்கள், இசை, சிற்பம் போன்றவற்றைக் கொண்டிருக்கிறது.
- சமூக அறிவியல் தத்துவம், வரலாறு, கல்வி போன்றவற்றைக் கொண்டிருக்கிறது.
- இயற்கை அறிவியல் இயற்பியல், வேதியியல், கணிதவியல், புவியியல் போன்றவற்றைக் கொண்டிருக்கிறது.
- வணிகவியல் வரவுசெலவு, வங்கி போன்றவற்றைக் கொண்டிருக்கிறது.
- மக்கள் தகவல் தொடர்பு செய்தித்தாள், காலவிதழ்கள், சுவரொட்டிகள், அறிவிப்புகள், விளம்பரங்கள் போன்றவற்றைக் கொண்டிருக்கிறது.
- மொழிப்பெயர்ப்பு தமிழில் மொழி பெயர்க்கப்பட்ட எல்லாப் விஷயங்களையும் உள்ளடக்கும்.

4.7. தரவு மாதிப்படுத்தலின் நெறிமுறை

தரவு ஒருவரின் தேவைக்கேற்ப சேகரிக்கப்பட்ட பொருள்களிலிருந்து (விஷயங்களிலிருந்து) பிரித்தெடுக்கப்படவேண்டும். விரிதரவின் கூடுதலான பிரதிநித்துவத்தை நிறுவுவதற்கு பல வழிகள் இருக்கின்றன. நாம் விரிதரவு மாதிரிப்படுத்தலின் வழிமுறைகளை வரையறை விளக்கம் செய்வதற்கு முன் நாம் ஆய விரும்புகிற மொழியின் வகையைத் தெளிவாக வரையறை விளக்கம் செய்யவேண்டும். முறையற்ற மாதிரிப்படுத்தும் உபாயம் (random sampling technique) ஒரு விரிதரவைத் திரிக்கப்பட்டதாக மற்றும் பிரதிநித்துவம் இல்லாததாகச் செய்வதிலிருந்து காக்கின்றது. இத்தரப்படுத்தப்பட்ட உபாயம் இயற்கை மற்றும் சமூக அறிவியில்களின் பல பரப்புகளில் பரவலாகப் பயன்படுத்தப்படுகின்றது. மற்றொரு வழி முழு நோக்கீட்டு நூல்களின் அகரவரிசை அட்டவணையைப் பயன்படுத்துவதாகும். British National Bbliography மற்றும் Willing's Press Guide என்பன LOB விரிதரவின் உருவாக்கத்திற்குப் பயன்படுத்தப்பட்டன. மற்றொருரு அணுகுமுறை ஒரு மாதிரிபடுத்தல் சட்டகத்தை வரையறை விளக்கம் செய்வதாகும். பிரவுன் விரிதரவின் திட்டவரைவாளர்கள் இதைப் பயன்படுத்தினார்கள். அவர்கள் ஒரு குறிப்பிட்ட ஆண்டில் வெளியிடப்பட்ட நூல்களையும் காலவிதழ்களையும் பயன்படுத்தினார்கள். ஒரு எழுத்து விரிதரவு செய்தித்தாள் அறிக்கை, காதல் கதைகள், சட்ட நியமங்கள், அறிவியல் எழுத்துரைகள், சமூக அறிவியல்கள், தொழில்நுட்ப அறிக்கைகள் போன்ற இனங்களிலிருந்து உருவாக்கப்பட்டலாம்.

4.8. தரவுகளை உள்ளீடு செய்யும் நெறிமுறை

தரவுகள் பின்வரும் வழிகளில் உள்ளீடு செய்யப்படலாம்:

மின் மூலத்திலிருந்து தரவு: இச்செயற்பாங்கில் மின்வடிவில் காணப்படுகின்ற செய்தித்தாள்கள், ஜர்னல்கள், மேகசின்கள், நூல்கள் என்பன உட்படுத்தப்படும்.

இணையவலையிலிருந்து தரவு: இது இணையப்பக்கம், இணைய தளம் மற்றும் தனிநபர் பக்கங்கள் என்பனவற்றிலிருந்து உரைகளை உட்படுத்தும்.

உரைகளைக் கணிப்பொறி படித்தறிதல் (Machine Reading of Text): இது ஒளி எழுத்துணரி ஒழுங்குமுறையால் (OCR) உரைகளைக் கணிப்பொறி படிக்கவியலும் வடிவில் மாற்றும். இம்முறையைப் பயன்படுத்தி அச்சடிக்கப்பட்ட விஷயப்பொருட்கள் விரிதரவாக விரைவில் மாற்றப்படுகின்றது.

மனித முயற்சியால் தரவை உள்ளீடுசெய்தல்: இது கணிப்பொறி விசைப்பலகையால் செய்யப்படுகின்றது. இது கையினால் எழுதப்பட்டவை, பேச்சு மொழியின் எழுத்துப்பெயர்ப்புகள், பழைய கையெழுத்துப்படிகள் என்பனவற்றிலிருந்து தரவைச் சேகரிப்பதற்கான நல்ல வழியாகும்.

தரவை உள்ளீடு செய்யும் செயன்மை மாதிரி நெறிமுறை அடிப்படையிலானது. நாம் ஒரு நூலிலிருந்து ஒவ்வொரு பத்து பக்கங்களுக்குப் பின்னரும் இரண்டு பக்கங்களை பயன்படுத்தவிலும். இது விரிதரவை உரைவடிவில் சேமிக்கப்பட்டுள்ள தரவின் நல்ல பிரதிநியாக மாற்றுகின்றது. ஒரு நூலில் வேறுபட்ட எழுத்தாளர்களால் எழுதப்பட்ட வேறுபட்ட விஷங்களை உள்ளடக்கிய பகுதிகள் இருந்தால் எல்லாப் பகுதிகளிலிருந்தும் இச்செயற்பாங்கால் சேகரிக்கப்படும் மாதிரிகள் சரியாகப் பிரதிநிதித்துவம் செய்யப்படும். தலைப்பு கோப்பு பதிவேடுகளைப் பாதுகாக்கவும் பதிப்புரிமைச் சிக்லைத் தீர்க்கவும் வேண்டி புத்தகத்தின் பெயர், படைப்பாளி/படைப்பாளிகளின் பெயர்கள், வெளியீட்டு ஆண்டு, பதிப்பு எண், வெளியிட்டவரின் பெயர், உள்ளீடாக எடுக்கப்பட்ட பக்கங்களின் எண் போன்ற எல்லா தகல்களையும் கொண்டிருக்கும்.

4.9. வன்பொருளின் தேவை

விரிதரவு உருவாக்கத்திற்குத் தேவையான கருவிகளில் எழுத்துரு அட்டை, மென்பொருள், எழுத்துரு ஆய்வி, காட்சித்திரை, அச்சு இயந்திரம், விசைப்பலகை, வன்தட்டு போன்றவை தேவை.

4.10. விரிதரவு கோப்புகளின் நிர்வாகம் (Management of Corpus File)

விரிதரவு நிர்வாகம் மிக கடினமான வேலை. இது வைத்திருப்பது, ஆய்வது, திரையிடுவது, பிற விரிதரவிலிருந்து தகவலை மீட்பது போன்ற வேலைகளை உள்ளடக்கும். ஒரு விரிதரவு உருவாகிவிட்டால் அது கணிப்பொறியில் சேமிக்கப்பட வேண்டும். அதைப் பாதுகாப்பதற்கும் விரிவாக்குவதற்கும்

திட்டங்கள் தேவை; தவறுகள் திருத்தப்பட வேண்டும். மாற்றங்கள் செய்யப்பட வேண்டும்; முன்னேற்றங்கள் நடைமுறைப்படுத்தப்பட வேண்டும்.

4.11. விரிதரவின் சீராக்கத்தின் நெறிமுறை (Methof of Corpus Sanitation)

தரவுகளை உள்ளீடு செய்த பின் அது திருத்தல்களுக்கும் மாற்றங்களுக்கும் உட்படுத்தப்பட வேண்டும். பொதுவாக நான்கு வகையான தவறுகள் நிகழலாம்:

1. எழுத்தை விட்டுவிடுதல் அல்லது நீக்குதல்
2. எழுத்தைச் சேர்த்தல் அல்லது திரும்பச் செய்தல்
3. எழுத்தைப் பதிலீடு செய்தல்
4. எழுத்தை இடம் மாற்றுதல் அல்லது இடம் பெயர்த்தல்

இவ்வெழுத்துப் பிழைகளை நீக்க நாம் விரிதரவை மனித முயற்சியால் பிரிசோதித்துத் திருத்தல் வேண்டும். அயல் மொழிச் சொற்கள், மேற்கோள்கள், கிளைமொழி வேறுபாடுகள் என்பன சரியாகத் தரப்பட வேண்டும். இவ்வாறு விரிதரவு சரிபார்க்கப்பட்டால் தரவை ஆய்வது எளிமையாக இருக்கும்.

4.12. பதிப்புரிமைச் சிக்கல் (Problem of Copyright)

நாம் எல்லாப் பதிப்புரிமை வைத்திருப்பவர்களிடமிருந்தும் (வெளியிட்டவர் மற்றும்/அல்லது படைப்பாளி, பேச்சு விஷயப்பொருள்களுக்கு எல்லாப் பேசுபவர்கள்) அனுமதி வாங்கிவைத்திருக்கவேண்டும். பதிப்புரிமைச் சட்டங்கள் சிக்சலானவை. எது சரியானது, எது தவறானது, எது சட்டபூர்வமானது எது சட்டத்திற்கு மாறானது என்பதைத் தீர்மானிப்பது கடினம். பதிப்புரிமை சிக்கல்கள் நாட்டிற்கு நாடு வேறுபடும். விஷயப்பொருட்கள் சொந்தப் பயன்பாட்டிற்கு என்றால் எவ்விதச் சிக்கலுமில்லை. அது வணிக நோக்கத்திற்காக நேரடியாகப் பயன்படுத்தாதது வரை எந்தவிதச் சிக்கலுமில்லை. விஷயப்பொருள்களைப் பயன்படுத்தி நாம் நாம் வணிகம் செய்ய புதிய கருவிகளையும் ஒழுங்குமுறைகளையும் உருவாக்கலாம். இந்நேர்வில் பதிப்புரிமை மீறப்படவில்லை. ஆனால் நேரடியான வணிகச் செயல்பாட்டிற்குச் சட்டபூர்வமான பதிப்புரிமை உள்ளவர்களிடமிருந்து முன் அனுமதி பெற்றிருக்கவேண்டும்.

பகுதி 5

உரை விரிதரவு ஆய்வு

5.0. அறிமுகம்

பல மொழிகளில் பெரிய மின்விரிதரவுகள் சேகரிக்கப்பட்டபின் அவற்றை ஆய்வு உத்திகளுக்கு உள்ளாக்கும் தேவை வரும். விரிதரவிலிருந்து மொழித் தரவை அணுகுவதற்கும் தேவையான தகவல்களை மீளப்பெறுவதற்கும் மக்கள் ஒழுங்குமுறைகளும் உபாயங்களும் உருவாக்கியுள்ளனர். இவ்வாய்வுக் கருவிகள் மொழி ஆய்வுக்கும் மொழித் தொழில் நுட்ப முன்னேற்றத்திற்கும் பயனுள்ளதாக அமைந்துள்ளன. பலவிதமான ஆய்வு உத்திகள் இருக்கின்றன. எடுத்துக்காட்டாக,

1. புள்ளியியல் ஆய்வி (Statistical Analyzer)
2. தொடரடைவு ஆய்வி (Concordencer)
3. சொல் சேர்ந்து வருகை ஆய்வி (Lexical Collocator)
4. முக்கியசொல் காணும் ஆய்வி (Key-word Finder)
5. அகராதி சொல் ஆய்வி (Lemmatizer)
6. உருபனியல் ஆய்வியும் உருவாக்கியும் (Marphological analyzer and Gernerator)
7. சொல் ஆய்வி (Word Processer)
8. சொல் வகைப்பாடு அடையாளப்படுத்தி (Parts-of-Speech Tagger)
9. விரிதரவு இடச்சுட்டி (Corpus Anotator)
10. பகுத்துக்குறிப்பான் (Parser)

ஆங்கிலம், பிரஞ்சு, ஜெர்மன் போன்ற மொழிகளுக்கு விரிதரவை ஆய பல மென்பொருள்கள் உருவாக்கப்பட்டுள்ளன. ஆனால் இந்தி, தமிழ் போன்ற இந்திய மொழிகளுக்கு அத்தகைய மென்பொருள்கள் இனிமேல் தான் உருவாக்கப்பட வேண்டும். இதற்கான முயற்சிகள் எடுக்கப்பட்டு வருகின்றன.

5.1. நிகழ்வெண் ஆய்வு

மொழியியல் நீண்ட காலமாகவே புள்ளியியலுடனும், கணக்கியலுடனும் தொடர்புடையது. கணித மொழியியல், கணினி மொழியியல், விரிதரவு மொழியியல், பயன்பாட்டு மொழியியல், குற்றாய்வு மொழியியல், நடை அளவியல் (Stylometrics) போன்றவை இயற்கை மொழி விரிதரவிலிருந்து பெறப்படும் வேறுபட்ட புள்ளியியல் மற்றும் அளவியல் தகவல்களை வேண்டுகிறது. மொழியின் வேறுபட்ட பண்புக்கூறுகளைக் குறித்த புள்ளியியல் தகவலின் போதுமாக அறிவு இல்லாமல் நாம் மொழித் தரவை கையாளுவதிலும் கண்டறிவதிலும் தவறுகள் செய்வோம். விரிதரவைப் பண்புசார் ஆய்வுகளுக்கும் (qualitative analysis) அளவுசார் ஆய்வுகளுக்கும் (quantitative analysis) உட்படுத்தலாம். பண்பு சார் ஆய்வு முந்தைய அமைப்பு மொழியியல் ஆய்வை குறிக்கும். மாறாக அளவுசார் ஆய்வு விரிதரவின் மேல் அளவு அடிப்படையிலான ஆய்வின் அடிப்படையில் கிடைக்கும் முழுமையான மற்றும் விளக்கமான வர்ணனையைத் தருவதை இலக்காக கொள்கிறது.

இவ்விலக்ககை அடைய பலவிதமான புள்ளியியல் அணுகுமுறை இருக்கின்றன. அவையாவன,

1. வர்ணனை புள்ளியியல் அணுகுமுறை
2. அறிதல் புள்ளியியல் அணுகுமுறை
3. மதிப்பீட்டுப் புள்ளியியல் அணுகுமுறை
4. பன்முக புள்ளியியல் அணுகுமுறை

5.2. சொல் வரிசைப்படுத்துதல் (Word Sorting)

அளவுசார் தரவில் நேரடியான அணுகுமுறை எண்ணிக்கை அடிப்படையிலான வரிசைப்படுத்தலாகும். எளிய நிகழ்வெண் எண்ணிக்கையிலிருந்து பெறப்படும் தகவலை நெடுங்கணக்கு வரிசையிலும் (Alphabetical Order) எண்ணிக்கை வரிசையிலும் (Numerical Order) வரிசைப்படுத்தலாம். இவற்றை மீண்டும் ஏறு வரிசையிலும் இறங்குவரிசையிலும் தரலாம். சொற்களின் நிகழ்வெண்பட்டியல் உரைகளுக்கு ஒரு குழும தடயங்களாகும். உரையை ஆயும் எவரும் ஒவ்வொரு சொல்லும் உரையில் எத்தனை முறை வருகிறது என்பதை அறிவதில் ஆர்வம் காட்டுவர். இப்பட்டியலைப் பரிசோதித்து நாம் ஒரு குறிப்பிட்ட உரையின் அமைப்பைப் பற்றி புரிந்துக்கொள்ளலாம். சொல்லில் மட்டுமின்றி எழுத்துக்கள், மரபுத்தொடர்கள், எச்சத்தொடர்கள் மற்றும் வாக்கியங்கள் இவற்றின் புள்ளி விவரங்களை நாம் விரிதரவிலிருந்து அறிந்துக் கொள்ளலாம். இது ஒரு மொழியின் மொழி அலகுகளையும் பண்புக்கூறுகளையும் புரிந்துக் கொள்ள பெரிதும் உதவும். எடுத்துக்காட்டாக, நடையியல் ஆய்வுக்கும் இத்தகைய புள்ளியியல் தகவல்கள் பெரிதும் கைகொடுக்கும்.

5.3. தொடரடைவு ஆய்வி (Concordencer)

தொடரடைவின் ஆய்வு விரிதரவில் சொற்களின் அகரவரிசை அட்டவணையை உருவாக்கப் பயன்படுகின்றது. இது ஒவ்வொரு சொற்களின் உரைசூழலுடன் கூடிய அவற்றின் நிகழ்வுகளின் சேகரிப்பாகும். ஒவ்வொரு சொல்லும் உரையில் அதன் வருகையிடம் குறிப்பிடப்பட்டு அட்டவணைப்படுத்தப்படும். உரையிலுள்ள பல முக்கியமான மொழி அமைப்பொழுங்குகளை அணுக உதவுவதால் இது மிக முக்கியமானதாகும். இது உள்ளுணர்வால் அறியவியாலாத தகல்களைத் தருகின்றது.

5.4. சொல் சேர்ந்து வருகை ஆய்வி (Lexical Collocator)

சொற்களின் சேர்ந்து வருகை உரைகளில் சொற்களின் பங்களிப்பையும் இடத்தையும் புரிந்துக் கொள்ள பெரிதும் உதவும். இவ்வாய்வு சொற்களில் எந்த இணைகளுக்கிடையில் கூடுதல் சேர்ந்து வருகை உறவு இருக்கின்றது என்பதை அறிய இயலும். மேலும் இரண்டு சொற்கள் சேர்ந்து வரும் நிகழ்வுத்தகைமையைத் தற்செயலான விளைவின் அடிப்டையிலான நிகழ்வுத் தகைமையுடன் ஒப்பிடும். ஒவ்வொரு சொல் இணைக்கும் ஒரு மதிப்பெண் தரப்படும்; கூடுதல் மிப்பெண் கூடுதல் சேர்ந்துவருகையைக் காட்டும். இது அகராதியிலும் தொழில்நுட்ப மொழிபெயர்ப்பிலும் பயன்படுத்த விரிதரவிலிருந்து பல்சொல் அலகுகளை பிரித்ததெடுக்க உதவும். இயந்திர மொழி பெயர்ப்புக்கும் பொருண்மை வேறுபாட்டிற்கும் இத்தகைய தகவல்கள் கைக் கொடுக்கும். இது அர்த்த வேறுபாடுகளை அடையாளம் காண ஒற்றுமையான சொற்களை குழும உதவுகின்றது.

எ.கா.

மாடிப்–படி – step

வாசற்-படி – step

புத்தகம் படி – read

சொன்னப்படி – manner

ஐந்து லிட்டர் படி – measure

ஒரே பொருண்மையைக் கொண்ட சொல்களுக்கு இடையேயுள்ள பயன்பாட்டு வேறுபாடுகளை வேறுபடுத்த உதவும்.

5.5. சூழலில் முக்கிய சொல் (Key Word in Context (KWIC))

தகவல் ஆய்வில் 'KWIC' (க்விக்) என்ற சொல் அடிக்கடி பயன்படுத்தப்படுகிறது. இது குறிப்பிட்ட சொற்களின் நேர்வுகளை அறிய உதவுகிறது. ஆய்வுக்கு உட்படும் சொல் ஒவ்வொரு வரியின் மத்தியில் கூடுதலான இடைவெளியுடன் தோன்றும். வேறுபட்ட நோக்கங்களுக்கு வேண்டி சூழலின் நீட்சி குறிப்பிடப்பட்டிருக்கும். இது மையத்திலுள்ள சொல்லின் இரண்டு பக்கங்களிலும் இரண்டு, மூன்று அல்லது நான்கு சொற்களின் சூழலைக் காட்டும். இதை N-gram (bi-gram, trigram, tetra–gram set) என்பர். இவ்வமைப்பொழுங்கு ஒருவரின் தேவை அடிப்படையில் மாறும். சொற்கள், தொடர்கள் மற்றும் எச்சத்தொடர்கள் இவற்றின் ஆய்வின் போது கூடுதலான சூழல் அவற்றைப் புரிந்துக் கொள்ளத் தேவைப்படும். கிவிக் என்பதை ஒரு உரையாக எண்ணுவது நல்லது; சொற்களின் நிகழ்வெண்ணை

மையச்சொல்லின் சூழலில் பரிசோதிக்கவேண்டும். எல்லாத் தகவல்களும் எப்போதும் தேவைப்படுவதில்லை; ஆனால் நாம் தேவைப்படும்போது தகவலைப் பயன்படுத்துகின்றோம்.

KWIC சொல்லால் பெருந்தரவை அணுகிய பின் நாம் மொழியியல் வர்ணணையில் பல்வேறுபட்ட நோக்கங்களை உருவாக்கலாம் மற்றும் இந்நோக்கங்களை நிறைவேற்ற வழிமுறைகளை வகுக்கலாம். KWIC சூழலில் முக்கியத்துவம், சேர்ந்து வரும் சொற்களின் பங்களிப்பு, சூழல்களில் சொற்களின் உண்மை நடத்தை, நேர்வின் உண்மைச் சூழ்நிலை மற்றும் சூழல் அடிப்படையிலான கட்டுப்பாடுகள் இவற்றைப் புரிந்துக் கொள்ள KWIC உதவும்.

5.6. குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதல் (Local Word Groupping)

குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதல் உரைகளில் சொற்கள் பயன்படுத்தப்படும் அமைப்பொழுங்கைப் பற்றிய தகவல்களை நமக்குத் தரும் மற்றொரு வகையான ஆய்வாகும். இது எங்கு சொல் நிரல் வாக்கியங்களின் பொருண்மைச் சுமையைத் தீர்மானிக்கின்றதோ அங்கும் எங்கு ஒரு உறுப்பின் தனிப்பட்ட பொருண்மைச் சுமை மற்றொரு உறுப்பின் வருகையால் பாதிக்கப்படுகிறதோ அங்கும் முக்கியத்துவம் வாய்ந்ததாகும். குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதல் (தொடர்கள் மற்றும் வாக்கியங்களின்) பகுத்துக்குறித்தலின் போது உறுப்புக்களின் செயல்பாட்டு நடத்தையை நேரிட தகவலைத் தருகிறது.

இது பயன்பாட்டில் வழக்கமற்ற ஆனால் தற்சுட்டு வடிவுகளுடன் தனிப்பட்ட உறவு கொண்ட வினைவடிவுகளின் (எகா. amuse onself, please oneself, lend onself, remind oneself) வகைமுறையைத் தீர்மானிக்கின்றது. இம்மாதிரியான அமைப்பொழுகின் அறிவு, மொழி கற்பவர்களுக்கு இடைப்பட்ட நிலையிலிருந்து உயர்ந்த நிலைக்குச் செல்ல முக்கியமாகும். குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதலைப் பயன்படுத்தி நாம் தமிழில் முற்று வினைகள், பெரும்பாலான எச்ச வினைகள் ளைக் கூடுதலாகத் தொடரும் என்றும் பெயர்கள் ஒட்டுகளாலும் பின்னுருபுகளாலும் தொடரப்படும் என்றும் அறிந்து கொள்ளலாம். குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதலால் கிடைக்கும் தகவல்கள் சொற்களை வினைக் குழுக்களாகவும் பெயர்க் குழுக்களாகவும் ஆய உதவும். குறிப்பிட்ட இடம் சார்ந்த சொற்களை குழுமுதலால் கிடைக்கும் தகவல் சொற்களின் குறிப்பிட்ட இடம் சார் சேர்க்கையின் காரணமாக வரும் சொல் மயக்கத்தைத் தீர்க்கின்றது. பொருண்மையின் நுண்மையான வேறுபாடுகள் சூழல்களில் சொற்களின் வருகைமுறையுடன் கூடிய உறுப்புக்களுக்கிடையே உள்ள அக உறவுகளால்

பெரும்பாலும் தரப்படுகின்றது. கூட்டுப் பெயர்களுக்கும் கூட்டு வினைகளுக்கும் சொற்களின் குறிப்பிட்ட சேர்க்கையால் உணர்த்தப்படும் பொருண்மையை தனிப்பட்ட சொற்களின் பொருண்மைகளிலிருந்து பெறவியலாது.

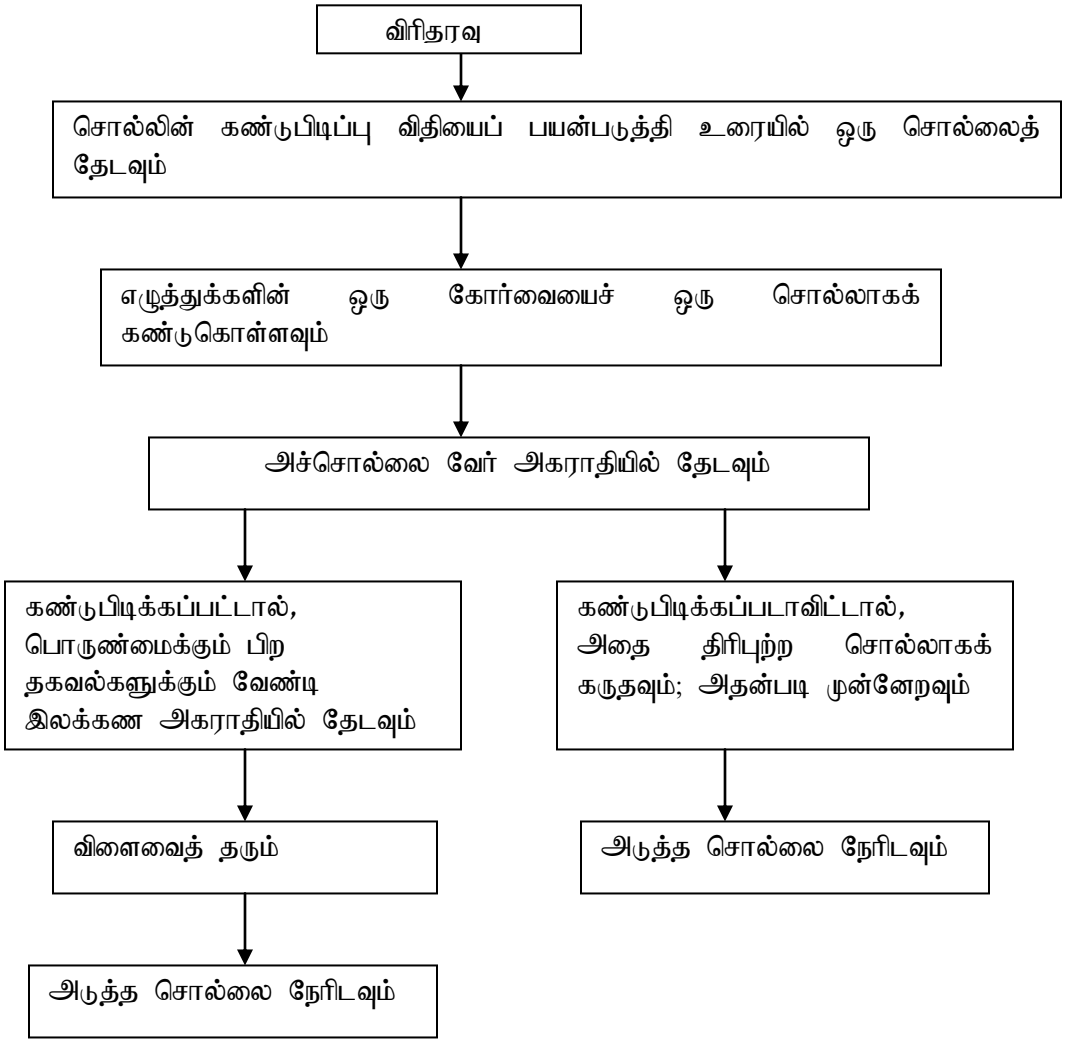
5.7. சொல் பகுப்பாய்வு (Word Processing)

சொல் பகுப்பாய்வு விரிதரவில் பயன்படுத்தப்படும் சொற்களைத் தானியக்கமாக ஆய்வதை உள்ளடக்கும். இதன் முக்கிய நோக்கம், ஒரு உரையின் பகுதிலிருந்து ஒரு சொல்லைக் கண்டுபிடித்து அதை அதன் பயன்பாட்டின் சூழலிருந்து பிரித்து, அதன் உருபொலியனியல் அமைப்பை ஆய்ந்து, அதன் மூலப்பொருண்மையை பெற்று, உரையில் அதன் தொடரியல் பங்களிப்பை விளக்குவது ஆகும்.

சொல் பகுப்பாய்விலிருந்து கிடைக்கும் தகவல் சொல் அர்த்த மயக்கத்தைத் தீர்ப்பதற்கும் (Word Sense Disambiguation (WSD) அகராதி உருவாக்குவதற்கும் பகுத்துக் குறிப்பதற்கும் மொழி கற்பதற்கும் மற்றும் இது போன்ற பலவற்றிற்கும் பயனுள்ளதாகும். ஆங்கிலம் போன்ற மொழிகளுக்குப் பல சொல் பகுப்பாய்வுகள் உருவாக்கப்பட்டு உள்ளன. தமிழ் போன்ற இந்திய மொழிகளுக்கும் இத்தகைய பகுப்பாய்வு உருவாக்கப்பட்டு பயன்படுத்தப்பட்டு வருகின்றன.

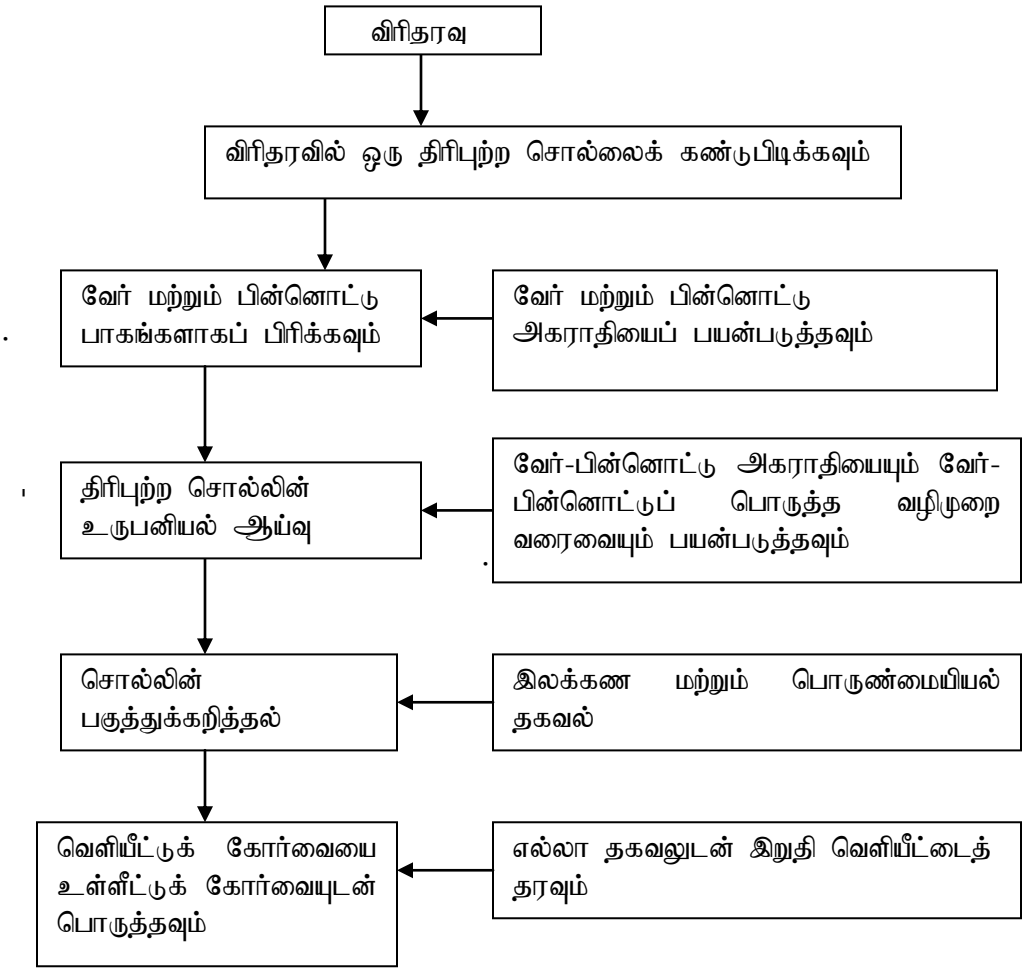
5.7.1. திரிபுறாத சொற்களின் ஆய்வு (Processing non-inflectecd forms)

பின்வரும் ஒழுக்குப்படம் திரிபுறாத சொற்களின் பகுப்பாய்வை விளக்கும்:


விரிதரவு
சொல்லின் கண்டுபிடிப்பு விதியைப் பயன்படுத்தி உரையில் ஒரு சொல்லைத் தேடவும்
எழுத்துக்களின் ஒரு கோர்வையைச் ஒரு சொல்லாகக் கண்டுகொள்ளவும்
அச்சொல்லை வேர் அகராதியில் தேடவும்
கண்டுபிடிக்கப்பட்டால், பொருண்மைக்கும் பிற தகவல்களுக்கும் வேண்டி இலக்கண அகராதியில் தேடவும்
கண்டுபிடிக்கப்படாவிட்டால், அதை திரிபுற்ற சொல்லாகக் கருதவும்; அதன்படி முன்னேறவும்
விளைவைத் தரும்
அடுத்த சொல்லை நேரிடவும்
அடுத்த சொல்லை நேரிடவும்

5.7.2. திரிபுற்ற சொற்களைப் பகுத்தாய்தல் (Processing Inflected Words)

பின்வரும் ஒழுக்குப்படம் திரிபுறாத சொற்களின் பகுப்பாய்வை விளக்கும்:


விரிதரவு
விரிதரவில் ஒரு திரிபுற்ற சொல்லைக் கண்டுபிடிக்கவும்
வேர் மற்றும் பின்னொட்டு பாகங்களாகப் பிரிக்கவும்
வேர் மற்றும் பின்னொட்டு அகராதியைப் பயன்படுத்தவும்
திரிபுற்ற சொல்லின் உருபனியல் ஆய்வு
வேர்-பின்னொட்டு அகராதியையும் வேர்-பின்னொட்டுப் பொருத்த வழிமுறை வரைவையும் பயன்படுத்தவும்
சொல்லின் பகுத்துக்கறித்தல்
இலக்கண மற்றும் பொருண்மையியல் தகவல்
வெளியீட்டுக் கோர்வையை உள்ளீட்டுக் கோர்வையுடன் பொருத்தவும்
எல்லா தகவலுடன் இறுதி வெளியீட்டைத் தரவும்

5.7.3. இரட்டை சொற்களை பகுத்தாய்தல் (Processing Double Words)

இரட்டைச் சொற்களின் ஆய்வு உறுப்புகள் அவற்றிற்கு இடையில் ஒரு இடைவெளியால் பிரிக்கப்பட்ட கூட்டுகள், இரட்டுற்ற சொற்கள், வேறுபடுத்தப்பட்ட சொற்கள் என்பனவற்றை உள்ளடக்கும். எல்லா வேறுபடுத்தப்பட்ட சொற்களும் பல்சொல் கோர்வைகளாகும்; அவை மிகத்திறமையுடைய ஆய்வு மற்றும் குறிப்புரை (annotation) வழி கையாளப்படவேண்டும். இரட்டை சொற்களை விரிதரவில் கண்டுபிடித்து ஆய்வது சிக்கலான செயல்முறையாகும். அகராதிகளில் பெரும்பாலும் இத்தகைய இரட்டை சொற்கள் தரப்பட்டிருக்கும். அத்தகைய சொற்களைப் பொருத்தமான அணுகுமுறையால் விரிதரில் அடையாளம் கண்டுக் கொள்ளலாம். ஆனால் அகராதியில் தரப்படாத இரட்டைச் சொற்களை அடையாளம் காண்பது சிக்கலான செயல் முறையாகும். சேர்ந்து வருகை நிகழ்வெண்கள் இதற்குக் கைகொடுக்கும். அடுத்துவருகின்ற சொற்களின் பொருண்மை இரட்டைச்

சொற்களின் பொருண்மையை உறுதிசெய்ய உதவுவதால் இரட்டை சொற்களை ஆய்தல் சொல் மட்டத்தில் பொருண்மை மயக்கம் தீர்க்க உதவும்.

5.8. அடையாயப்படுத்துகை

குறிப்பிட்ட பண்புக்கூறுகளைக் குறிப்பிட வேண்டி சொற்களுக்குத் தனிப்பட்ட குறியங்களை ஒட்டுவதை உள்ளடக்கும் மொழிக் குறிப்புரையின் (linguistic annotation) சில வகைகள் குறிப்புரை, குறியங்கள் என்று அறியப்படுவதற்குப் பதிலாக அடையாளப்படுத்துதல் என்று அறியப்படுகின்றது.

5.8.1. சொல் வகைப்பாட்டு அடையாளப்படுத்துகை (Part of Speech Tagging)

சொல்வகை அடையாளப்படுத்தும் திட்டம் ஒரு சொல்லை வாக்கியத்தில் அதன் சொல்வகைப்பாட்டுடன் அடையாளப்படுத்தும். இது மூன்று நிலைகளில் செய்யப்படுகின்றது: முன்திருத்தல், தானியங்கு அடையாளம் தருகை, மற்றும் மனிதமுயற்சிசார் பின்திருத்தல்.

அடையாளப்படுத்தாத வாக்கியம் (Untagged sentence)

எ.கா.

A more to stop Mr. Gaitskell from nominting any more labour life peers is to be made at a meetiong of labour MPs tomorrow.

அடையாளப்படுத்தப்பட்ட மாதிரிவடிவம் 1 (Tagging Sample 1)

^a_ AT move_ NN to_ TO stop _VB \0Mr_ NPT Gaitskell _NP from _IN nominating _ABG any_DTI. more_AP labour_NN life_NN peers_NNS is_BEZ to-TO be_BE made–VBN at_IN a_AT meeting –NN of_IN labour _NN tomorrow _NR._

5.8.2. இலக்கண அடையாளப்படுத்தல்

சொல்வகை அடையாளப்படுத்தலில் இரண்டாவது மட்டம் இலக்கண அடையாளப்படுத்தல் எனப்படும். இது விரிதரவில் உள்ள ஒவ்வொரு சொல்லுக்கும் இலக்கணத் தகவலைத் தரும் கூடுதல் விரிவான குறிப்புரை ஒழுங்குமுறையாகும். ஆங்கிலத்திற்கு சில இலக்கண அடையாளப்படுத்தல்

கருவிகள் உருவாக்கப்பட்டுள்ளன. இலக்கண அடையாளப்படுத்தல் தட்டம் இறுதி மனிதமுயற்சிசார் மதிப்பீட்டிற்கும் திருத்தலுக்கும் திட்டவரைவு செய்யப்பட்டுள்ளது.

5.8.3. சொல் அர்த்த அடையாளப்படுத்தல் (Word Sense Tagging)

சொல் அர்த்த அடையாளப்படுத்தல் சொல் அடையாளப்படுத்தப்பட்ட மற்றும் இலக்கணம் அடையாளப்படுத்தப்பட்ட உள்ளீட்டு உரையின் சொற்களை ஏற்கும். தானியங்கு அடையாளப்படுத்தலுக்குப் பின் ஒவ்வொரு சொல்லும் சரியான பொருண்மை வகுப்பாக்கத்தை உறதி செய்ய வேண்டி மனித முயற்சிசார் பின்திருத்தல் செய்யப்படும்.

5.9. தலைச்சொல்லாக்கம் (Lemmatisation)

தலைச்சொல்லாக்க செயல்பாடு விரிதரவில் பயன்படுத்தப்பட்டுள்ள சொற்களின் சொல்வகைப்பாட்டை கண்டுபிடிப்பதுடனும் அவற்றுடன் தொடர்புடைய சொல்லன்களாகக் (lexeme) குறைப்பதுடனும் தொடர்புடையதாகும். எடுத்துக்காட்டாக, படித்தான், படிக்கிறான் என்ற சொல்லிலிருந்து 'படி' என குறைத்து தலைச்சொல்லாக்கம் செய்தல்.

இது சொல்லின் எல்லா மாற்று வடிவங்களையும் உள்ளீடு செய்யாமல் ஒரு தலைச் சொல்லின் மாற்று வடிவங்களைப் பிரித்தெடுக்கவும் பரிசோதிக்கவும் ஆய்வாளர்களை அனுமதிக்கிறது. இது மொழி கற்றலுக்குப் பயன் உள்ளதாக அமைகிறது. இங்கு கற்பவர் ஒரு தலைப்புச் சொல்லின் மொத்த சாத்தியமான எண்ணிக்கையில் பயிற்சி பெறுகிறார்கள். இது எந்தச் சொற்கள் திரிபுறுகின்றன, எத்தனை தடவைத் திரிபுறுகின்றன எவ்வாறு திரிபுறுகின்றன என்பதை அறிவதற்குப் பயன் உள்ளதாக அமையும். எடுத்துக்காட்டாக, Brown Corpus-இல் ஒரு பகுதி சொல் மற்றும் இலக்கண தகவலுடன் கூடிய சொற்களின் தலைப்பாக்கம் செய்யப்பட்ட வடிவங்களைக் கொண்டிருக்கிறது.

5.10. குறிப்புரை செய்தல் (Annotation)

தனியான உரைகளுடன் மட்டுமின்றி விரிதரவு குறிப்புரை செய்தல் என்று அழைக்கப்படுகின்ற கூடுதல் மொழியியல் தரவுகளைக் கொண்டிருக்கும். இத்தகவல்கள் சொல்வகை குறிப்புரைசெய்தல் (Parts of Speech Annotation), மீக்கூறு அடையாளப்படுத்தல் (Prosodic Annotation), பொருண்மை

அடையாளப்படுத்தல் (Semantic Annotation), முன் வருகிளவி அடையாளப்படுத்தல் (Anaphoric Annotation), கருத்தாடல் அடையாளப்படுத்தல் (Discourse Annotation) என வேறுபடும்.

அடையாளப்படுத்தப்பட்ட விரிதரவு ஆய்வுக்கு மிகப் பயனுள்ளதாக அமையும். இலக்கணத்திற்காக அடையாளப்படுத்தப்பட்ட விரிதரவு குறிப்புரை செய்யப்பட்ட விரிதரவுகளில் மிகப் பொதுவானதாகும்; இதில் சொற்களுக்குச் சொல் வகுப்புகள் தரப்பட்டிருக்கும். Brown Corpus, LOB Corpus, BNC என்பன இலக்கணத்திற்காகக் குறிப்புரை செய்யப்பட்ட விரிதரவு ஆகும். LLC மீக்கூறுகளுக்காகக் குறிப்புரை செய்யப்பட்ட விரிதரவு ஆகும். Susanne Corpus தொடரியலுக்காகக் குறிப்புரைசெய்யப்பட்ட விரிதரவு ஆகும். இந்திய மொழிகளின் விரிதரவு அடையாளப்படுத்தலின் செயல் இன்னும் சரியாகத் தொடங்கபடவில்லை.

5.10.1. சொல் வகைப்பாட்டு குறிப்புரை செய்தல் (Part of Speech Annotation)

சொல்வகைப்பாட்டு குறிப்புரை செய்வதன் நோக்கம் ஒரு உரையில் உள்ள ஒவ்வொரு சொல் அலகுக்கும் அதன் சொல் வகைப்பாட்டை காட்டும் ஒரு குறியனைத் தருவதாகும். இது பெருந்தரவிலிருந்து தரவை மீட்டெடுப்பதில் தனித்தன்மையைக் காட்டுகிறது மற்றும் தொடரியல் பகுத்துக்குறித்தலுக்கு உதவுகிறது. இது பொருண்மைக்கள குறிப்புரை செய்தலுக்கும் உதவுகிறது. இது ஒற்றுமையான எழுத்துகளை வேறுபடுத்துவதற்கு நம்மை அனுமதிக்கிறது.

5.10.2. முன்வருகிளவி குறிப்புரை செய்தல் (Anaphoric Annotation)

முன்வருகிளவி குறிப்புரைத் திட்டத்தில் எல்லா மாற்றுப்பெயர்களும் பெயர்த்தொடர்களும் பற்றாசு (Cochesion) என்பதன் விரிவான சட்டகத்திற்குள் ஒரே குறிபடுபொருளைக் குறிப்பதற்காக குறிப்புரை செய்யப்படுகின்றது. இங்கு முன்வருகிளவி வேறுபட்ட வகைகளுக்காகப் பட்டியலிட்ப்பட்டு வரிசைப்படுத்தப்படுகின்றன. இம்முன்வருகிளவி குறிப்புரை செய்யும் திட்டம் மாற்றுப்பெயர் தீர்வு போன்ற ஆய்வுகளுக்கும் பரிசோதனை செய்யும் இயங்குமுறைகளுக்கும் பயன்படுத்தப்படுகிறது. இது உரையைப் புரிந்து கொள்வதற்கும் இயந்திர மொழதுபெயர்ப்புக்கும் முக்கயமானதாகும்.

5.10.3. மீக்கூறு குறிப்புரை செய்தல் (Prosotic Annotation)

மீக்கூறு குறிப்புரை செய்தலில் இசையோட்ட ஒழுங்கமைப்புகள், அழுத்தம் மற்றும் விட்டிசை என்பன பேச்சில் அடையாளப்படுத்தப்படும். இது மிகச் சிரமமான குறிப்புரை ஆகும்; ஏனென்றால் மீக்கூறு பிற மொழி மட்டங்களை விட இயல்பில் தன்னுணர்ச்சி அடிப்படையானது. இது பயிற்சி பெற்ற காதால் கவனமாகக் கேட்கப்படுதலைக் வேண்டும். Lancoster / IBM Spoken English Corpus மீக்கூறுக்காகக் குறிப்புரை செய்யப்பட்ட விரிதரவாகும். இவற்றில் அழுத்தப்பட்ட அசைகள் சுதந்திரமான மற்றும் சுதந்திரமற்ற இசைமை இயக்கத்துடன் வேறுபட்ட குறியீடுகளால் அடையாளப்படுத்தப்பட்டுள்ளது. சுற்றுப்புற ஒலியழுத்தமுள்ள அசைகளின் சுரக் குறிப்பிலிருந்து இசையைச் ஊகிக்க இயலும் அழுத்தமில்லா அசைகள் அடையாளப்படுத்தப்படாமல் விடப்பட்டுள்ளன.

5.10.4. பொருண்மை குறிப்புரை செய்தல் (Sementic Annotation)

பொருண்மையைக் குறிப்புரை செய்தல் உரையிலுள்ள சொற்களின் பொருண்மைப் பண்புக்கூறுகளை (குறிப்பாகச் சொல் அர்த்தங்களைக் குறிப்புரை) அல்லது சொற்களுக்கு இடையே உள்ள பொருண்மை உறவுகளைக் (குறிப்பிட்ட செயலின் செயலிகள் பாதிக்கப்படுவர் போன்றவற்றை) குறிப்புரை செய்யும். எடுத்துக்காட்டாக, அவன் பழம் சாப்பிட்டான் என்ற வாக்கியம் பின்வருமாறு பொருண்மைக் குறிப்புரை செய்யப்படும்:

அவன் <படர்க்கை, ஆண்பால் ஒருமை, மாற்றுபெயர்>
பழம் <தாவரத்தின் பூவிலிருந்து விளையும் இனிப்பான பொருள்>
சாப்பிட்டான் <உணவை உட்கொண்டான்>

அவன் <எழுவாய்> பழம் <செயப்படுபொருள்> சாப்பிட்டான் <பயனிலை>

எந்தப் பொருண்மை பண்புக்கூறுகள் அடையாளப்படுத்தப்பட வேண்டும் என்பது ஒருங்கிணைந்த கருத்து இல்லை. சிலர் Rogest thesaurus-ல் உள்ள பொதுமையான பொருண்மை வகைப்பாடுகளை முன்வைக்கின்றனர். அம்மாதிரியான குறிப்புரை செய்யும் திட்டம் சொற்களின் மூடப்பட்ட வகுப்புகளுக்கும் திறந்த வகுப்புகளுக்கும் நடைமுறைப்படுத்தவேண்டி வடிவமைக்கப்பட்டுள்ளது.

5.10.5. கருத்தாடல் குறிப்புரை செய்தல் (Discoursal Annotation)

கருத்தாடல் குறிப்புரை செய்தலில் ஒரு விரிதரவு உரை மற்றும் கருத்தாடல் மட்டங்களில் குறிப்புரை செய்யப்படுகிறது மற்றும் மொழியியல் ஆய்வில் பயன்படுத்தப்படுகிறது. இவ்வகைக் குறிப்புரை கருத்தாடல் ஆய்வில் முக்கிய பங்களிப்பை செய்தாலும் பரவலாகப் பயன்படுத்தப்படவில்லை; ஏனென்றால் மொழி வகைப்பாடுகள் சூழல் சார்ந்தவை மற்றும் அவற்றை உரைகளில் அடையாளம் காணுதல் பிற மொழியியல் நடப்புகளைக் காட்டிலும் விவாதத்திற்குக் கூடுதல் முல காரணமாக அமையும். வயதுக்குவந்தவர்களின் பேச்சு போக்கைக் கவனிக்க வேண்டி சிலர் London-Lund Spoken Corpus-ஐ 16 கருத்தாடல் அடையாளங்களால் குறிப்புரை செய்துள்ளனர்.

5.11. பகுத்துக்குறித்தல் (Parsing)

பகுத்துக் குறித்தல் என்பது இலக்கண அடிப்படையில் உரைகளின் தானியக்க ஆய்வுடன் தொடர்புடையது. தொழில் நுட்ப அடிப்படையில் இது ஒரு உரைக்குத் தொடரியல் அமைப்பைத் தரும் செயற்பாங்கை குறிப்பிடுகிறது. இது பெரும்பாலும் உரையில் அடிப்படை உருபன்-தொடரியல் வகைப்பாடுகள் கண்டுபிடிக்கப்பட்ட பின் நடைமுறைபடுத்தப்படும். வேறுபட்ட இலக்கணங்களின் (சார்பு இலக்கணம், சூழல் கட்டில்லா இலக்கணம், ஒழுங்குமுறையுள்ள செயல்பாட்டு இலக்கணம், நீட்சிசெய்யப்பட்ட ஒட்டு இருக்கணம் போன்றன) அடிப்படையில் பகுத்துக்குறித்தல் இந்த உருபன்-தொடரியல் வகைப்பாடுகளை ஒன்றுடன் ஒன்று உயர்ந்த மட்ட தொடரியல் உறவுகளில் கொண்டுவருகிறது. வாக்கிய நிலை பகுத்துக்குறித்தல் சொல்மட்ட பகுப்பாய்விலிருந்து பெறப்பட்ட தகவலைப் பயன்படுத்தி தானியக்கச் சூழல் அடிப்படையிலான மற்றும் சூழல் சுதந்திரமான தொடரியல் ஆய்வை உட்படுத்தும். ஒரு பகுத்து குறிக்கப்பட்ட விரிதரவு பகுத்துக்குறித்தலில் கிளைபடங்களைப் பயன்படுத்துவதால் இது கிளைவங்கி (Tree Bank) என்று அழைக்கப்படும். கிளை அமைப்பின் காட்சிப்படம் மிக அரிதாகவே விரிதரவு அடையாளப்படுத்தலில் காணப்படும். பொதுவாக ஒத்த தகவல் புலக்குறிப்பு செய்யப்பட்ட அடைப்புக்குறிகளின் குழுமங்களைப் பயன்படுத்தி உருபடுத்தம் செய்யப்படும். எடுத்துக்காட்டு, Pearl sat on a chair என்பது கிளை வங்கியில் பின்வருமாறு தோன்றும்:

[S[NP Pearl_ NP1 NP] [VP sat_VVD [PP on_II [NP a_AT1 chair_ NN1 NP] PP] VP]S]

இதில் உருபனியல்-தொடரியல் தகவல்கள் கீழ்க்கோடுகளால் சொற்களுடன் இணைக்கப்படடுள்ளது; உறுப்புகள் தொடர்களின் தொடக்கத்திலும் இறுதியிலும் முறையே திறக்கும் மற்றும் அடைக்கும் செவ்வக அமைப்புக் குறிகளால் குறிப்புரை செய்யப்பட்டுள்ளது.

எல்லாப் பகுத்துக்குறித்தல் ஒழுங்கு முறைகளும் ஒன்றல்ல. முக்கியமான வேறுபாடுகள் பின்வருவன:

1. ஒரு ஒழுங்குமுறை பயன்படுத்தும் உறுப்பு வகைகளின் எண்ணிக்கை
2. இந்த உறுப்பு வகைகள் ஒன்றுடன் ஒன்று இணைவதற்கு அனுமதிக்கப்படும் வழி

இவ்வேறுபாடுகள் இருப்பினும் பெரும்பாலான பகுத்துக்குறித்தல் திட்டங்கள் சூழல் கட்டுப்பாடு இல்லாத தொடரமைப்பு இலக்கணத்தின் அடிப்படையில் அமைந்தது.

இவ்வொழுங்குமுறைக்குள் முழு பகுத்துக் குறித்தல் (Full Parsing) திட்டம் வாக்கிய அமைப்பின் விரிவான ஆய்வைத் தருவதை நோக்கமாக கொண்டது. குறுகிய பகுத்துக்குறித்தல் திட்டம் (Skeleton Parsing Scheme) தொடரியல் உறுப்பு வகைகளில் குறைந்த நிலையிலான வேறுபடுத்தப்பட்ட குழுமத்தைப் பயன்படுத்துகிறது மற்றும் சில உறுப்பு வகைகளின் அக அமைப்பை விட்டுவிடுகிறது. பகுத்துக்குறித்தல் பெரும்பாலும் மனித ஆய்வால் பின்திருத்தம் செய்யப்படுகிறது; ஏனென்றால் தனியங்கி பகுத்துக்குறித்தல் சொல் வகை அடையாளப்படுத்தலை விட வெற்றி விகித்தில் குறைந்தது. முழு மினித இயக்கப் பகுத்து குறித்தலின் குறைபாடு, விரிதரவு பகுத்துக்குறித்தல் அல்லது திருத்தல் என்பனவற்றில் ஈடுப்படும் ஆய்வாளரின் ஒழுங்கின்மையாகும். இதை ஈடுகட்ட விரிவான வழிமுறைகள் தரப்படுகின்றன. இருப்பினும் பல்பொருள்கோள் சாத்தியமாகும் போது மயக்கம் ஏற்படலாம். கிளை வங்கிகள் இயற்கை மொழிகளுக்கு அதன் அமைப்பின் சொல் மட்டம், தொடர்மட்டம், வாக்கியமட்டம், செயல்பாடு-பங்கெடுப்பாளர் அமைப்பு மட்டம் என்ற வேறுபட்ட அமைப்புகளின் மட்டங்களில் குறிப்புரைகளைத் தரும் மொழி மூலவளமாக அமைகிறது. கிளைவங்கிகள் இயற்கை மொழி ஆய்வில் தரவு-இயக்க அணுகுமுறைகள், மனித மொழி தொழில் நுட்பங்கள், இலக்கணப் பிரித்தெடுப்பு மற்றும் பொதுவான மொழியியல் ஆய்வுகள் இவற்றின் முன்னேற்றதத்திற்கு முக்கியமானதாகும்.

பகுதி 6

மொழித் தொழில் நுட்பத்தில் விரிதரவு

6.0 அறிமுகம்

மொழித் தொழில் நுட்பத்தில் விரிதரவின் பங்களிப்பு இன்றைய காலகட்டத்தில் மிக சிறப்பாகவும் பரவலாகவும் பேசப்படும் மற்றும் பயன்படுத்தப்படும் நெறிமுறையாகும். மொழி புரிந்துகொள்தல், பேச்சு புரிந்துக் கொள்ளுதல், உரை மீட்பு மற்றும் புரிந்துகொள்ளுதல், உரைகளிலிருந்து தகவல்களை மீட்டல், ஒலிவழி எழுத்து புரிந்துக்கொள்ளுதல் இயந்திரமொழிபெயர்ப்பு போன்றவற்றை உள்ளடக்கும். கணிப்பொறி அடிப்படையிலான மொழியியல் தொழில்நுட்பத்தில் ஆர்வம் கூடி வருகிறது. இருப்பினும் மனித மொழியில் உட்படும் கலவைத்தன்மைகள் மற்றும் சிக்கல்கள் காரணமாக மொழி ஆய்வுக்காக திட்டமிடப்படும் கணினி வழியமைப்பு முறைகள் வேறுபட்ட மொழித் தரவுகளின் மிகக் கூடுதலான அளவை வேண்டும். இதன் காரணமாக அளவீட்டு மொழியியல் என்று அழைக்கப்படும் விரிதரவு மொழியியல் பயன்பாட்டு மொழியியலின் புதிய கிளையாக உருவாகியுள்ளது. எல்லா மொழிகளிலும் விரிதரவு உருவாக்கப்பட்டு அவற்றை மொழியின் பல நிலைகளிலும் அடையாளப்படுத்தும் முயற்சி நடைப்பெற்று வருகிறது. இத்தகைய விரிதரவு மொழிக் கருவிகள் தயாரிப்பதற்கு மிகப்பயனுள்ளதாக அமைகிறது. இந்நெறிமுறை மொழித் தொழில் நுட்பமாக வளர்ந்துள்ளது.

6.1 மொழித்தொழில் நுட்பத்தில் பெருந்தரவின் பங்கு

வேறுபட்ட வகையான பெருந்தரவு உருவாக்கத்தால் மொழித்தொழில் நுட்ப ஆய்வும் பயன்பாட்டிலும் பெருந்தரவின் முக்கியத்துவம் பெருகியுள்ளது. மொழி ஆய்வுக்கான கருவிகள் மற்றும் ஒழுங்கு முறைகள் உருவாக்குவதில் மொழி விரிதரவு பெரிதும் பங்களிப்பு செய்கிறது. பயன்பாட்டு அடிப்படையிலான தொழில் நுட்பச் செயலுக்கு விரிதரவு உயர்ந்த விளைவைத் தருகிறது. பொதுவாக நாம் விரிதரவின் பயன்பாட்டை இயந்திரத்தால் கட்டுப்படுத்தப்பட்ட கருவிகளை திட்டமிடுவதற்கும் தானியங்கு கருவிகளைப் பரிசோதிப்பதற்கும் பயிற்சி தரவும் சிறப்பான மூலவளமாக காண்கிறோம். விரிதரவிலிருந்து பெறப்பட்ட மொழிப் பண்புக்கூறுகளின் நிகழ்வெண்ணிக்கை மொழி கற்போருக்கு நூல்களைத் திட்டமிடவும், ஒளிவழி எழுத்துக்களைப் புரிந்துகொள்ளும் ஒழுங்குமுறைகளை **(OCR)** உருவாக்குவதற்கும் தானியங்கு எழுத்துப்பிழைத் திருத்திகளை உருவாக்குவதற்கும் பயனுள்ளதாக இருக்கிறது. குறிப்புரை செய்யப்பட்ட மற்றும் குறிப்புரை செய்யப்படாத விரிதரவுகள் இயந்திரமொழிபெயர்ப்பு ஒழுங்குமுறையை திட்டமிடவும் இயந்திரத்தால் படிக்கப்பெறும் அகராதிகளைத் திட்டமிடவும் (Machine Readable Dictionaries) MRDS)) உருபனியல் ஆய்விகள், சொல் ஆய்விகள், வாக்கியப் பகுத்துக்குறிப்பான்கள் என்பனவற்றை உருவாக்கவும் பயனுள்ளதாக அமைகிறது. விரிதரவைப் பயன்படுத்தி உருவாக்கப்பட்ட பல கணினி கருவிகளும் ஒழுங்குமுறைகளும் மொழியைப் பயன்படுத்துபவர்களுக்கும் ஆய்வாளர்களுக்கும் எழுத்தாளர்களுக்கும் கல்வியாளர்களுக்கும் ஆசிரியர்களுக்கும் மாணவர்களுக்கும் அறிஞர்களுக்கும் அச்சிட்டு வெளியிடுபவர்களுக்கும் மொழி கற்பவர்களுக்கும் மற்றும் பிறருக்கும் மிகுந்த பயனுள்ளதாக அமைகிறது. பொதுவாக விரிதரவின் முக்கியதுவத்தை மொழித் தொழில்நுட்பத்தின் நான்கு விரிந்த களங்களாகப் பகுக்கலாம்; இது அவற்றின் இயல்பு மற்றும் பயன்பாட்டு நோக்கு அடிப்படையில் செய்யப்படும்:

- அறிவின் மூலவளமாக விரிதரவு
- மொழித் தொழில்நுட்ப கருவிகளையும் ஒழுங்குமுறைகளையும் திட்டமிட மூலவளமாக விரிதரவு
- மொழிபெயர்ப்புக்கு உதவும் ஒழுங்குமுறைகளின் மூலவளமாக விரிதரவு
- மனித இயந்திர இடைமுக (இடையூடக) ஒழுங்குமுறைகளின் மூலவளமாக விரிதரவு

6.2. அறிவின் மூலவளமாக பெருந்தரவு

அறிவு மூலவளமாக விரிதரவு பின்வரும் செயல்பாடுகளுக்கு பயன்படுகிறது:

1. பன்மொழி நூலகங்களை உருவாக்குதல்

2. மொழி கற்பவர்களுக்குப் பாடநூல்களைத் திட்டமிடல்
3. (அச்சிட்ட மற்றும் மின்வடிவ) ஒருமொழி அகராதிகளை உருவாக்குதல்
4. (அச்சிட்ட மற்றும் மின் வடிவ) இருமொழி அகராதிகளை உருவாக்குதல்
5. (அச்சிட்ட மற்றும் மின் வடிவ) பன்மொழி அகராதிகளை உருவாக்குதல்
6. (அச்சிட்ட மற்றும் மின்வடிவ) ஒருமொழிச் சொற்களஞ்சியங்களை (பொருட்புல அகராதிகளை) உருவாக்குதல்
7. பல்வேறு விதமான (அச்சிட்ட மற்றும் மின் வடிவ) நோக்கீட்டு பொருள்கள் உருவாக்குதல்
8. இயந்திரத்தால் படிக்கவியலும் அகராதிகளை உருவாக்குதல்
9. பன்மொழி சொல் மூலவளங்களை உருவாக்குதல்
10. மின் அகராதிகளை உருவாக்குதல்

6.3 மொழித் தொழில் நுட்பக் கருவிகளைத் திட்டமிடுதல்

மொழித் தொழில் நுட்பக்கருவிகளி பினவருவனவற்றை உள்ளடக்கும்:

1. சொல்லாய்வு ஒழுங்குமுறை
2. எழுத்துப்பிழைத்திருத்தும் ஒழுங்குமுறை
3. உரை நேர்செய்யும் ஒழுங்குமுறை
4. உருபனியல் பகுப்பாய்வு ஒழுங்குமுறை
5. வாக்கிய பகுத்துக்குறிப்பு ஒழுங்குமுறை
6. நிகழ்வெண் கணக்கிடும் ஒழுங்குமுறை
7. சொல் தேடும் இயந்திரம்
8. உரை சுருக்கும் ஒழுங்குமுறை
9. உரை அடையாளப்படுத்தும் ஒழுங்குமுறை
10. தகவல் மீட்கும் ஒழுங்குமுறை
11. தொடரடைவு ஒழுங்குமுறை
12. சொற்பொருள் மயக்கம் நீக்கும் ஒழுங்குமுறை
13. சொல் வலை திட்டமிடல்
14. பொருண்மை வலை திட்டமிடல்
15. சொல் வகைப்பாடு அடையாளப்படுத்தும் ஒழுங்குமுறை

16. ஓரிடம்சார் சொற்களை குழுமும் ஒழுங்குமுறை

6.4 மொழிபெயர்ப்புக்கு உதவும் ஒழுங்குமுறைகளுக்கு மூலவளமாகப் பெருந்தரவு

இது பின்வருவனவற்றை உள்ளடக்கும்:

1. மொழி மூலவளம்பெறும் ஒழுங்குமுறை
2. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை
3. பன்மொழித் தகவல் பெறும் ஒழுங்குமுறை
4. மொழி கடந்த தகவல் மீட்பு ஒழுங்குமுறை

6.5 மனித-இயந்திர இடைமுக ஒழுங்குமுறைகளுக்கு மூலவளமாக பெருந்தரவு

இது பின்வருவனவற்றை உள்ளடக்கும்:

1. ஒலியால் வருடி எழுத்துக்களைப் புரிந்துக்கொள்ளும் ஒழுங்குமுறைகள்
2. குரலைப் புரிந்துக்கொள்ளும் ஒழுங்குமுறைகள்
3. உரையிலிருந்து பேச்சு ஒழுங்குமுறைகள்
4. வலை (இணையத்தளம்) அடிப்படையிலான கற்றல் ஒழுங்குமுறைகள்
5. கேள்வி-பதில் ஒழுங்குமுறைகள்
6. கணினி உதவியுடன் கட்டளைகள்
7. கணினியின் உதயுடன் மொழிக் கல்வி
8. உரை உருவாக்கம்

6.6 பேச்சுத் தொழில்நுட்பத்தில் விரிதரவு

இது பின்வருவனவற்றை உள்ளடக்கும்:

1. பேச்சுத் தொழில்நுட்பத்திற்குப் பொதுவான சட்டகத்தின் உருவாக்கம்
2. கிளைமொழிகளில் ஒலியியல், சொல்லியியல் மற்றும் உச்சரிப்பியல் வேறுபாடுகள்
3. தானியங்கு பேச்சு புரிந்துக்கொள்ளுதல் செயல்பாடு

4. தானியங்கு பேச்சு உருவாக்கும் ஒழுங்குமுறைகள்
5. தானியங்கு பகுப்பாய்வு ஒழுங்குமுறைகள்
6. தானியங்கு பேசுபவரை அடையாளம் காணும் ஒழுங்குமுறைகள்
7. பேச்சுக் குறைபாடுகளைச் சரிசெய்தல்

இயல்/பகுதி - 7

மொழியியலின் விரிதரவு பயன்பாடு

7.0 அறிமுகம்

விரிதரவானது முக்கிய மொழியியல் ஆய்விலும் பகுப்பாய்விலும் பயன்பாட்டிலும் முதன்மையான மூலப்பொருளாக மாறியுள்ளது. இது மொழியியலில் ஏற்பட்டுள்ள அறிவியல் பார்வையையும் தொழில் நுட்ப பார்வையும் பிரதிபலிக்கிறது.

7.2 அகராதியியலில் விரிதரவு (Corpus In Lexicography)

தரப்படுத்தப்பட்ட அகராதிகள் ஒரு மொழியில் பயன்படுத்தப்படும் சொற்களுக்கு உருபனியல், மூலப்பொருளியியல், பொருண்மையியல் மற்றும் ஒலியியல் தகவல்களைத் தருகின்றது. அவை அகராதியை உருவாக்குபவரின் அறிவைப் பிரதிபலிப்பதாக அமைக்கிறதேயன்றி மொழி பயன்பாட்டாளர்களின் சொல் பற்றிய அறிவை வெளிப்படுத்துவதாக அமையவில்லை. விரிதரவு இச்சிக்கலுக்கு விடை தருகிறது. விரிதரவு பல வகைப்பட்ட மொழி வகைகளின் மற்றும் மொழி நடைகளின் மாதிரித் தரவுகளைக் கொண்டிருப்பதன் காரணமாக சொற்களின் பலவிதமான பயன்பாட்டு அடிப்படையிலான பொருண்மைகளைப் பொருத்தமான எடுத்துக்காட்டுகளுடன் பெற உதவுகின்றது. ஒரு சொல்லுக்கு எத்தனை பொருள்கள் பட்டியலிடப்பட வேண்டும் என்பதை தீர்மானிப்பதற்குப் விரிதரவு கைகொடுக்கின்றது. விரிதரவில் உள்ள சொல் குறித்த தரவுகளிலிருந்து நாம் சரியான பொருண்மையைத் தேர்ந்தெடுக்க இயலுகின்றது. ஒரு சொல்லின் பொருண்மை குறித்த நமது தவறான பார்வையைத் மாற்றிக்கொள்ள விரிதரவு உதவுகிறது. சொற்களின் பொருண்மை மாற்றத்தையும் புதிய பொருண்மைகளையும் அறிய இயல்வதுடன் மொழியில் பயன்படுத்தப்படுகின்ற புதிய சொற்களையும் வழக்கிழந்த சொற்களையும் அறிந்து கொள்ள இயலுகின்றது. விரிதரவில் பெறப்படுகின்ற சொற்கள் குறித்த புள்ளியியல் விவரங்கள் அடிப்படையில் பல விதமான அகராதிகளை உருவாக்க இயலும்.

எடுத்துக்காட்டாக, ஆரம்ப வகுப்பு மாணவர்களுக்கென்றும் நடுத்தர பள்ளி மாணவர்களுக்கென்றும், உயர்நிலை மாணவர்களுக்கென்றும், கல்லூரி மாணவர்களுக்கென்றும் தனித்தனியான அகராதிகளைப் விரிதரவின் அடிப்படையில் அமைத்துக் கொள்ளலாம். பொதுவான பயன்பாட்டாளர்களுக்கும் அகராதி உருவாக்க விரிதரவு கைக்கொடுக்கும். ஆங்கிலத்தில் BNC and Bank of English என்ற தரவின் அடிப்படையில் உருவாக்கப்பட்ட The Collin's COBUIID Dictionary ஆங்கிலத்தில் உருவாக்கப்பட்ட விரிதரவு அடிப்படையிலான மிகவும் பயன்பாட்டில் உள்ள அகராதியாகும்.

7.3 சொல்லியியலில் உள்ள விரிதரவு (Corpus In Lexicology)

சொல்லியியல் சொல்லைப் பற்றிய மிக ஆர்வமும் பயனுள்ள பரப்பெல்லையை உள்ளடக்குகிறது. இது சொற்களின் பொருண்மை மீட்பையும் சொற்களின் பொருண்மை மாற்றத்தையும் சொற்களின் வேறுபாடுகளையும் கால அடிப்படையில் சொல்வடிவங்களின் மாற்றத்தையும் பல நூற்றாண்டுகளில் சொற்றொகை வளர்ச்சியையும் புதிய சொல் ஆக்கங்களையும் சொல் இழப்பையும் சொல் கடன் வாங்கலையும் சொற்களின் அமைப்பு மற்றும் மூலப்பொருள் ஆய்வையும் உள்ளடக்கும். சொல்லியியலில் விரிதரவின் பயன்பாடு இந்திய மொழிகளுக்கு கூறத்தக்க அளவு முன்னேற்றம் அடையவில்லை. அதனால் ஆங்கிலம் போன்ற மொழிகளுக்கு இவ்வாய்வுகளம் பெரிதும் விரிவடைந்துள்ளது. எடுத்துக்காட்டாக, The Helsinki Corpus of English Text மற்றும் Mechigan Early Modern English Materials Corpus along with Oxford English Dictionary என்பன அருவப் பெயர்களின் பொருண்மையில் ஏற்படும் மாற்றங்களின் திசைகளையும் இயங்குமுறையையும் அறிய உதவுகின்றது. எடுத்துக்காட்டாக, Toronto Corpus of Middle English என்பது இலக்கியவயமான, உருவகவயமான மற்றும் ஆகுப்பெயர்வயமான அர்த்தங்களை மீட்டுருவாக்கம் செய்யப் பயன்பட்டது. இவ்வாறு பல சொல்லியியல் ஆய்வுகள் மொழியியல் ஆய்வுக்கு உறுதுணையாக நிற்கின்றன.

7.4 இலக்கண உருவாக்கத்தில் விரிதரவின் பங்கு (Corpus In Grammar Writing)

விரிதரவு இலக்கணங்களை எழுதுவதற்குப் பெரிதும் பங்களிப்பு செய்கின்து. ஆரம்ப கால கட்டங்களில் இலக்கணமானது இலக்கண ஆசிரியரின் உள்ளறிவின் அடிப்படையில் எழுதப்பட்டது. அப்படிப்பட்ட இலக்கணங்கள் விதிமுறைகளைக் கட்டளையிடுபவைகளாக அமைந்ததேயன்றி மொழியின் இலக்கண அமைப்பை விளக்குவதாக அமையவில்லை. ஆரம்ப கால மொழியியல் அடிப்படையிலான இலக்கணங்கள் கூட மொழித் தரவின் அடிப்படையில் இலக்கண விளக்கத்தை நோக்கமாகக் கொண்டு அமைந்ததேயன்றி அது ஒரு மொழியின் இலக்கணத்தின் எல்லா பரிமாணங்களையும் விளக்குவதாக

அமையவில்லை. இவ்விலக்கணம் பல குறைபாடுகளை நீக்குவதற்குத் துணை புரிகிறது. விரிதரவின் அடிப்படையில் எழுதப்பட்ட இலக்கணங்கள், எழுதப்படுகிற இலக்கணங்கள் விரிவானதாகவும் மொழியியல் அமைப்பை முழுவதுமாக வெளிப்படுத்துவதாக அமைகின்றன. விரிதரவு மொழியைச் சரியாக பிரதிபலிக்கும் இலக்கணங்களை உருவாக்க பயன்படுவதால் இத்தகைய விரிதரவு சார்ந்த இலக்கணம் மொழி பகுப்பாய்வுக்கும் உருவாக்கத்திற்கும் மொழிபெயர்ப்பிற்கும் மொழி கற்பித்தலுக்கும் பெரிதும் உதவும்.

7.5 பொருண்மையியல் ஆய்வில் விரிதரவு (Corpus In Semantic Study)

விரிதரவு அடிப்படையிலான மொழி ஆய்வு மொழி அலகுகளின் மற்றும் உரைப் பகுதிகளின் பொருண்மையியல் ஆய்வுக்கு ஒரு அறிவியல் அணுகுமுறையை அமைக்கிறது. இது ஒரு சொல்லின் பொருண்மை அது நிகழும் சூழலிலிருந்து ஆக்கப்படுகிறது என்று கூறுகின்றது. ஒரு சொல்லின் பொருளோ அல்லது ஒரு உரையின் பகுதியோ அல்லது ஒரு உரையின் பகுதியின் பொருளோ அது வரும் முந்தைய நிகழ்வின் வரலாற்றை உள்ளடக்கும். இது அச்சொல்லைப் பற்றி அச்சொல்லின் பொருண்மைப்பற்றி மற்றும் பகுதிகளின் பொருண்மைப் பற்றிய எல்லா தகவல்களையும் உள்ளடக்கும். சொற்களின் பொருண்மை அவை வரும் சூழல்களிலிருந்து பிரித்து எடுக்கப்படும். இத்தகைய செயல்பாட்டிற்கு விரிதரவு பெரிதும் உதவுகிறது. விரிதரவு ஒரு சொல் வரும் பல்வேறு வகையிலான சூழல்களை கொண்டிருப்பதால் விரிதரவை பயன்படுத்தி ஒரு சொல்லின் பொருண்மைகளைத் துல்லியமாக பட்டியலிடவியலும். சொற்களின் பொருண்மை சார் வேறுபாடுகள் அவை பயன்படுத்தப்படும் உருபனியல், தொடரியல், இசையோட்டம் மற்றும் மரபு தொடர் சட்டகங்களில் வெளிப்படுத்தப்படும் கூர்ந்து பார்க்கத்தக்க சூழல்களுடன் தொடர்புடையது என்று விரிதரவின் ஆய்வு காட்டுகின்றது. இதுபோல் கூட்டுச் சொற்கள், பல்சொல் அலகுககள், சேர்ந்துவரும் சொற்கள், குழுமத் தொடர்கள் என்பனவற்றின் பொருண்மையியல் ஆய்வுக்கும் புரிந்துகொள்ளலுக்கும் விரிதரவு அடிப்படையிலான சூழல் தகவல் தேவை. விரிதரவு அடிப்படையில் கண்டுபிடிக்கப்பட்டுள்ள அல்லது கண்டுபிடிக்கப்படும் வேறுபட்ட அர்த்தங்களின் எண்ணிக்கை சாதாரண அகராதியில் காணப்படும் அர்த்தங்களின் எண்ணிக்கைகளைக் காட்டிலும் மிக கூடுதலாக இருப்பதை காணலாம்.

சொற்களின் பல்பொருண்மையைச் சரியாகப் புரிந்துகொள்வதற்கு விரிதரவு மிக பயனுள்ளதாக அமைகிறது. சொற்களின் பல்பொருண்மை விரிதரவிலிருந்து எடுக்கப்பட்டால் அவை மிகச் செம்மையாக அமையும். இயற்கை மொழி ஆய்வில் மிகச் சிக்கலான பிரச்சனை சொற்களில் பொருண்மை மயக்கம் தீர்க்கும் சிக்கலாகும். குறிப்பாக இயந்திர மொழி பெயர்ப்புக்குச் சொற்களின் பொருண்மை மயக்கம்

தீர்ப்பது முக்கியமான செயல்பாடாகும். புஸ்தெஜொவ்ஸ்கியின் ஆக்கமுறை அகராதி சொற்களின் பொருளைச் சூழல்களிலிருந்து ஆக்குவதற்கான விதிகளை ஆய்கின்றது. விரிதரவு ஆய்வு இதற்கு பெரிதும் கைக்கொடுக்கும். சொற்களின் உருவக பொருளும் உருவாக அர்த்தமும் ஆகபெயர் அர்த்தமும் சூழல்களிலிருந்து பெற இயலுவதற்கு விரிதரவு பெரிதும் உதவுகின்றது.

7.6 மொழி கற்றலில் விரிதரவு (Corpus In Language Learning)

விரிதரவானது மொழி கற்பித்தலில் பெரிதும் பயன்படுகிறது. பொதுவாக விரிதரவிலிருந்து எடுக்கப்பட்ட எடுத்துக்காட்டுகள் ஒரு மொழியைக் கற்பவர்களுக்கு அவர்கள் தங்களுடைய உண்மையான வாழ்க்கைச் சூழல்களில் நேரிடும் மொழி வகையில் (வாக்கியங்கள், தொடர்கள், சொற்கள் போன்றவை) பயிற்சியளிக்கப் பயன்படுகின்றது. கற்பித்தலுக்கு மட்டுமல்லாமல் கற்பித்தலில் பயன்படுத்தப்படும் கற்றல் பொருட்களை உருவாக்கவும் விரிதரவு பெரிதும் பயன்படுகிறது.

பாடநூல் மூலம் மொழியைக்கற்றலுக்கும் தாய்மொழி பேசுபவர் சூழலுக்கு தகுந்தவாறு மொழியைப் பயன்படுத்துவதற்கும் இடையில் கூறத்தக்க வேறுபாடுகள் இருக்கின்றன என்பதை ஆய்வுகள் காட்டுகின்றன. பாடநூல்கள் சிலர் தேர்வு செய்யப்பட்ட மொழி அமைப்புகளைத் தான் கற்பவருக்குத் தருகின்றன. உண்மையான மொழி பயன்பாட்டுச் சூழலில் பயன்படுத்தப்படும் மொழியின் அமைப்பை அவை கற்றுதருவதில்லை. கற்றல் பொருட்கள் விரிதரவில் கிடைக்கப்பெறும் உண்மையான மொழிப் பயன்பாட்டின் அடிப்படையில் அமைந்த மொழிக் கூறுகளைக் கற்பவர்களுக்கு அறிமுகப்படுத்தும் படி அமைய வேண்டும். பர்லோ (Barlow, 1996) கூறுவது போல் விரிதரவுகள் கற்பவர் பெறவேண்டிய மொழியின் அமைப்பொழுங்கின் பரப்பை வெளிப்படுத்துவதுடன் அவற்றின் நிகழ்வெண்களையும் வெளிப்படுத்த வேண்டும். இது கற்றல் பொருள்களை உருவாக்குவதற்கும் பாடத் திட்டத்தை உருவாக்குவதற்கும் முக்கியமானதாகும். வாக்கியங்களில் மரபுத்தொடர்களைப் பயன்படுத்துதைக் கட்டுப்படுத்தும் கொள்கைகள் போன்ற மொழிப்பயன்பாட்டின் பிற பல்வேறுபட்ட மொழிக் கூறுகளையும் கற்பவர் புரிந்துக் கொள்ள விரிதரவு துணை செய்கின்றது. உள்ளறிவு அடிப்படையிலான மொழி கற்பித்தல் தவறான மொழிப் பேறுக்கு வழி செய்யும் ஆனால் விரிதரவு அடிப்படையிலான மொழி கற்பித்தல் நம்பத் தகுந்ததும் சிறப்பானதுமாக அமைகிறது என்று ஆய்வுகள் வெளிப்படுத்துகின்றன. கணிப்பொறி உதவியுடன் செய்யப்படும் 'இடையீட்டு மொழி கற்றலுக்கு' (interactive language learning) விரிதரவு பெரிதும் பயன்படுகிறது. International corpus of Leaner English வேறுபட்ட நாடுகளிலிருந்து

வரும் ஆங்கிலத்தை அயல் மொழியாக கற்றுக் கொண்டவர்களால் உருவாக்கப்பட்ட எழுத்தாண்மைகளிலிருந்து எடுக்கப்பட்டவைகளைக் கொண்டிருக்கிறது.

7.7 கிளைமொழி ஆய்வில் விரிதரவு (Corpus In Dialect Study)

விரிதரவு பல்வேறு இடங்களில் பேசப்படும் ஒரு மொழியின் வேறுபாடுகளை ஆய்வதற்குப் பெரிதும் துணைப்புரிகிறது. இதற்குக் கிளை மொழிகளின் விரிதரவு மிகவும் முக்கியமானதாகும். கிளைமொழியியலார்கள் எவ்வாறு மொழியானது இடத்திற்கிடம் வேறுபடுகிறது என்பதைப் பற்றி அறிய ஆர்வமுயைவராய் இருக்கின்றனர். முன்னர் மரபாகச் சேகரிக்கப்பட்ட கிளைமொழிகளின் தரவுகள் கிளைமொழிகளின் வேறுபாட்டை சரியாகப் பிரதிபலிப்பதாக அமையவில்லை. கிளை மொழிகளின் விரிதரவு கிளைமொழிகளின் அமைப்பை விளக்கும் தரவுகளைக் கொண்டிருப்பதால் இது மிகவும் சிறப்பாக அமைகிறது. கிளை மொழி மீதான முந்தைய ஆய்வுகள் திருப்திகரமாக அமையவில்லை. கிளைமொழிகளின் ஒலியனியல், உருபனியல், தொடரியல் அமைப்புகளை ஆய்ந்து ஒப்பிடவும் மீட்டுருவாக்கம் செய்யவும் கிளை மொழி அகராதியை உருவாக்கவும் கிளைமொழி விரிதரவு தான் நம்பத் தகுந்ததாக அமையும்.

கிளைமொழி விரிதரவு கிளைமொழி ஆய்வுகளுக்கு மிக பயனுள்ளதாக அமைந்துள்ளது. கிளைமொழிகள் எவ்வாறு ஒன்றிலிருந்து ஒன்று ஒலியனியல் அடிப்படையிலும் உருபனியல் அடிப்படையிலும் தொடரியல் அடிப்படையிலும் பொருண்மையியல் அடிப்படையிலும் வேறுபடுக்கின்றன என்றும் அவை எவ்வாறு நிலைபேறு மொழியிலிருந்து வேறுபடுகின்றன என்றும் கிளைமொழி விரிதரவு மூலம் அறியலாம். இந்திய மொழி சூழலைப் பொறுத்த வரையில் கிளைமொழி விரிதரவு மிகவும் பயனுள்ளதாகும்.

7.8 சமுதாய மொழியியலின் விரிதரவு (Coupus In Sociolinguistics)

சமுதாய மொழியியலில் விரிதரவின் பயன்பாடு மிக வலுவானதாக அமையும். விரிதரவுகள் சமுதாய அடிப்படையில் மொழிகள் எவ்வாறு வேறுபடுகின்றன என்பதையும் ஒரு சமுதாயத்தைச் சார்ந்தவரின் சமூகமாக்கத்தை அறிந்து கொள்ளவும் பயன்படும். சமுதாய மொழியியலில் பெரும்பாலான ஆய்வுகள் மொழிக்கும் பாலுக்கும் இடையே உள்ள ஊடாட்டத்தைக் கண்டுபிடிப்பதற்கான சொல்லியல் ஆய்வுகளுடன் தொடர்புடையதாக இருக்கின்றது. இவ்வாய்வுகள் குறிப்பிட்ட ஆய்வு சார்ந்த தரவுமையத்தின் எல்லைக்குட்பட்ட சேகரிப்பின் அடிப்படையில் அமைந்ததால் அவை அளவு

பரிசோதனைக்கும் சீரான முன்மாதிரிக்கும் உட்படுத்தப்படவில்லை. இக்குறைபாடுகளை நீக்குவதற்காக விரிதரவு இயல்பான தரவுகளின் ஏரானமான பிரதிநிதித்துவ மாதிரிகளைத் தருகின்றது; இதை அளவீட்டிற்கும் சீரான மாரிதிக்கும் உட்படுத்தலாம். விரிதரவு பேசுபவர்களைப் பற்றியும் பயன்படுத்துபவர்களைப் பற்றியும் படைப்பாளர்களைப் பற்றியும் பல்வேறு விதமான சமூக மொழியியல் தகவல்களைக் கொண்டிருப்பதால் இது சமுதாய மொழியியல் ஆய்வுகளுக்கு திறமையான மூலவளமாக அமைகிறது. எடுத்துக்காட்டாக, BNC விரிதரின் பெரும்பாகம் பேசுபவரின் வயது, பால், சமூக வகுப்பு, எழுதுபவரின் வயது, பால், இருப்பிடம் போன்ற சில சமுதாய மொழியியல் மாறிகளுக்கு அடையாளப்படுத்தப்பட்டிருப்பதால் சமுதாய மொழியியல் ஆய்வுக்கும் விசாரணைக்கும் பயன்படுகிறது

7.9 உளமொழியியலில் விரிதரவு (Corpus in Psycholinguistics)

தற்போதைய உளமொழியியல் ஆய்வுகளிலும் பயன்பாடுகளிலும் விரிதரவு பரிசோதனைக் கூடப் பரிசோதனைகளால் உருவாக்கப்பட்ட தரவுக்கான மூல வளமாகப் பயன்படுத்தப்படுகிறது. பல்வேறுபட்ட சொல் அலகுகளின் நிகழ்வெண் சொல் புரிந்துகொள்ளலை உள்ளடக்கும் பல புலனறிச் செயற்பாங்குகளுக்குள் முக்கிய அளவீடாக இருப்பதால் முறையாக மாதிரிப்படுத்தப்பட்ட விரிதரவு, சொல் பயன்பாட்டின் நிகழ்வெண்ணைப் பற்றிய உறுதியான மற்றும் நம்பத்தகுந்த தகவல்களைத் தருவதற்குப் பயனுள்ளதாகக் கருதப்படுகின்றது. விரிதரவில் பிரதிபலிக்கப்படும் மொழியின் பயன்பாட்டைக் கூர்ந்துநோக்கி உளமொழியியல் மனித மூளையால் மொழி எவ்வாறு ஆயப்படுகின்றது என்பதைப்பற்றிய கருதுகோளின் உருவாக்கத்திற்குப் பங்களிப்பு செய்யும். உளமொழியியலில் விரிதரவின் தவிர்க்கவியலாத பயன்பாடு இயல்பான உரையாடல்களில் ஏற்படும் பேச்சுத் தவறுகளை பரிசோதிப்பதிலால் அறியப்படும். எடுத்துக்காட்டாக, London-Lund Speech Corpus ஆங்கிலத்தில் இயல்பான உரையாடலில் ஏற்படும் பேச்சுத் தவறுகளை ஆய்வதற்கு நம்பகமான தரவு மையமாகும். மொழிக் குறைபாடுகளைப் பற்றி ஆய்வதற்கும் விரிதரவு மிக பயனுள்ளதாக அமையும். இவ்விரிதரவு மொழிக்குறைபாடுடைய பேசுபவர்களின் மாதிரித் தரவுகளைக் கொண்டிருக்கவேண்டும்.

7.9 நடையியலில் விரிதரவு (Corpus Stylistics)

விரிதரவு வேறுபட்ட மொழி வகைகள், பாடக் களங்கள், எழுத்தாளர்கள், தகவல் தொடர்புகள் போன்றவற்றின் உரை மாதிரிகளின் மிகுதியான திரட்டலைக் கொண்டிருப்பதால் விரிதரவு நடையியல் ஆய்வுகளுக்குப் பெரிதும் பயனுள்ளதாக அமைகிறது. ஆய்வாளர்கள் தனிப்பட்ட உரை வகைகளிலும் நடையியல் அடிப்படையில் ஆர்வம் காட்டுகின்றனர். சிலர் ஒரு குறிப்பிட்ட நாட்டைச் சார்ந்த

எழுத்தாளர்களின் நடையை அறிவதிலும் ஆர்வம் காட்டுகின்றனர். சிலர் நடைகளை ஒப்பீடுவதிலும் ஆர்வம் காட்டுகின்றனர். விரிதரவு இத்தகைய ஆய்வுக்கெல்லாம் வேண்டிய விரிவான தரவைக் கொண்டிருக்கிறது.

பகுதி 8

இயந்திர மொழிபெயர்ப்பில் விரிதரவு

8.0 அறிமுகம்

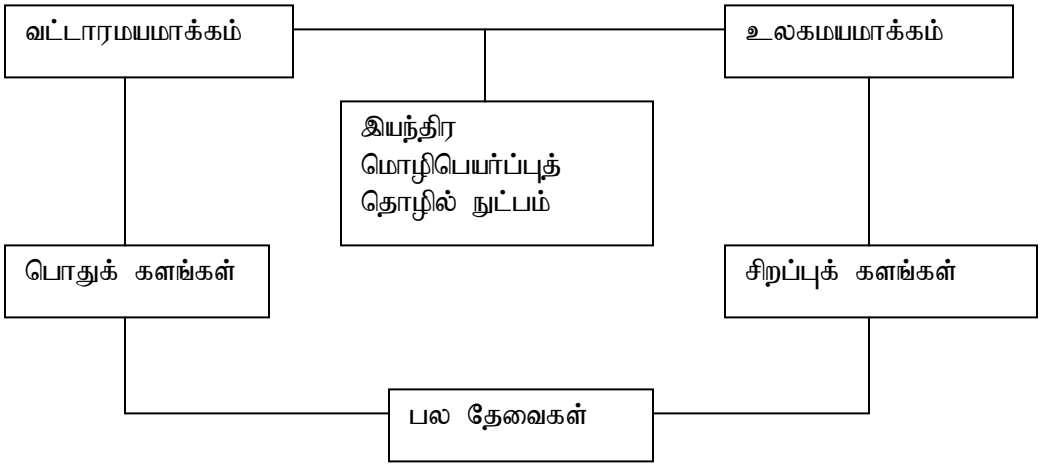
இயந்திர மொழிபெயர்ப்பு செயற்கை அறிவு நட்பத்தின் சிக்கலான ஒரு பகுதியாகும். இது மூல மொழியின் வாக்கியங்களிலிருந்து, இலக்குமொழி வாக்கியங்களை உருவாக்குவதற்கான ஒழுங்கு முறையை உருவாக்க வேண்டும். இலக்கு மொழியின் வெளியீடு உயர்ந்த தரமுடையதாக அமையத் தேவையில்லை; ஓரளவுக்குத் தரம் வாய்ந்ததாய் இருந்தால் போதும். இயந்திர மொழிபெயர்ப்பு மொழிகளுக்கிடையே கருத்துப்பரிமாற்றத்திற்கும் தகவல் பரிமாற்றத்திற்கும் மொழிகளைக் கடந்த தகவல் மீட்புக்கும் மொழி கற்றலுக்கும் முக்கியத் தொழில் நுட்பமாகும். பல விதமான பயன்பாட்டின் சாத்தியம் சாரணமாக இயந்திர மொழிபெயர்ப்பு மின் வணிகம் தகவல் வட்டாரமாக்கம் பல்மொழி ஆவணப்படுத்தல் மற்றும் தகவல் தொழில் நுட்பம் போன்றவற்றில் மிகப் பயனுள்ள தொழில் நுட்பமாக விளங்குகிறது. நாம் இயந்திர மொழிபெயர்ப்புத் தொழில் நுட்பத்தில் விரிதரவின் பங்களிப்பைப் பற்றிக் கூறுவதற்கு முன் பின்வரும் கேள்விக்குத் திருப்திகரமான விடை தர வேண்டும்: மனிதர்கள் மொழி பெயர்ப்பைச்

செய்யும்போது ஏன் இயந்திர மொழி பெயர்ப்பை நாம் ஏன் உருவாக்க வேண்டும்? இக்கேள்விக்கான பதில் எளிமையானது அல்ல. ஹுட்சின் (Hutchin's 1986) பின்வரும் கதரணங்களைத் தருகிறார்:

1. பல்வேறுபட்ட அறிவியல் தொழில் நுட்ப ஆவணங்களை உலகிலுள்ள தொழில் நுட்பவியலாருக்கும் அறிவியலாருக்கும் உடனடியாகத் தர வேண்டுமானால் இயந்திர மொழிபெயர்ப்பு மிகவும் அவசியமாகும்.
2. மனித மொழி பெயர்ப்பாளர்கள் இல்லாத சூழல்களில் இயந்திர மொழிபெயர்ப்பு மிகவும் பயனுள்ளதாக அமையும்.
3. இயந்திர மொழிபெயர்ப்பு மொழிகளைக் கடந்த ஆவண மாற்றங்கள் மூலம் அக நிலையானக் கூட்டுறவை மேம்படுத்த இயலும்.
4. இயந்திர மொழிபெயர்ப்பு பல்வேறுபட்ட அறிவியல், தொழில்நுட்பம், விவசாயம் மற்றும் மருத்துவ தகவல்களை ஏழ்மையான மற்றும் முன்னேறுகின்ற நாடுகளுக்கு விரைவான, எளிதான மற்றும் மலிவான பரிமாற்றம் அடிப்படையில் மொழித் தடைகளை நீக்கி அக நிலையானக் கூட்டுறவையும் அமைதியையும் மேம்படுத்த இயலும்.
5. மொழிப் பெயர்ப்பு இராணுவ நோக்கங்களுக்கும் தொழில்நுட்ப ஆய்வு மற்றும் வணிக நோக்கங்களுக்கும் மிகவும் பயனுள்ளது.

1980-களில் ஹுட்சின் இயந்திர மொழிப்பெயர்ப்புக்குச் சாதகமாகக் கூறிய போது இருந்த சூழல் கடந்த ஆண்டுகளில் மிகவும் மாறியுள்ளது. இந்நூற்றாண்டின் தொடக்கத்தில் மக்களின் வாழ்க்கை நடத்தை காரணமாக தாய்மொழி மூலம் தகவலைப் பெறுவது விதியாக மாறியுள்ளது. மேலும் படிப்பறிவின் வளர்ச்சி, செய்தி மற்றும் தகவலின் உலகமயமாக்கம், நாடுகளைக் கடந்த பன்மொழிய விரிதரவு, இனங்களுக்கிடையே மொழித் தனித்தன்மையின் அவசியம் மற்றும் பிற காரணிகள் கணினி அறிவியலாரையும் மொழியியலாரையும் இயந்திர மொழிப்பெயர்ப்பு ஒழுங்குமுறையை உருவாக்க இணைந்து பணியாற்றும் சூழலுக்குத் தள்ளியுள்ளது.

இக்காரணிகளின் பங்களிப்பைப் பின்வரும் வரைபடத்தின் மூலம் சுருக்கமாகக் கூறலாம்:



இங்கு அனுபவவாத வரிதரவு அடிப்படையிலான மொழிபெயர்ப்பு முன்மொழியப்படுகின்றது. இது மொழிபெயர்ப்பில் ஈடுபடுத்தப்பட்டுள்ள மொழிகளின் உரைகளிலிருந்து உருவாக்கப்பட்ட மொழி பெயர்க்கப்பட்ட விரிதரவுகளின் பகுப்பாய்விலிருந்து பெறப்பட்டத் தகவல்களின் மற்றும் எடுத்துக்காட்டுகளின் அடிப்படையில் அமைந்தது. இங்கு இயந்திர உதவியால் செய்யப்படும் மனித மொழிபெயர்ப்பிற்கு, மொழிபெயர்ப்பு விரிதரவு பயன்படுத்தப்பட விவாதிக்கப்படுகின்றது; எல்லைப்படுத்தப்பட்ட மற்றும் பொது களங்களில் இவ்வொழுங்குமுறையின் பயன்பாடு அழுந்தக்கூறப்படுகின்றது. தகவல் பரிமாற்றத்தின் உலக உயர்வுக்கும் வட்டாரவாக்கத்திற்கும் மொழிபெயர்ப்பு ஒழுங்குமுறையின் தேவை விளக்கப்பட்டு எதிர்கால முன்னேற்றத்திற்கு வேண்டிய வழிமுறைகள் முன்வைக்கப்படுகின்றன. மொழி பெயர்ப்பு வரலாற்றின் மற்றும் தோல்வியில் இன்று கற்றுக்கொண்டவைகளின் அடிப்படையில் இயந்திர மொழிபெயர்ப்பு அமைக்கப்பட்ட ஆலோசனை கூறப்பட்டுள்ளது.

8.1. நோக்கம் (The Objective)

கடந்த 50 ஆண்டுகளில் இயந்திர மொழிபெயர்ப்பு தொழில் நுட்பங்கள் மற்றும் பயன்பாடு இவற்றின் அடிப்படையில் கூறத்தக்க அளவு வளர்ந்துள்ளது. இன்றைய இணைய தள கால கட்டத்தில் இணைய தளத்தில் தகவல்கள் பெரிய அளவில் பல மொழிகளில் உள்ளன. அவற்றை எல்லா மொழி பேசுவோரும் பகிர்ந்து கொள்ள வேண்டுமானால் அவை ஒரு மொழியிலிருந்து மற்றொரு மொழிக்கு

உடனடியாக மொழி மாற்றம் செய்யப்பட வேண்டும். இதை இயந்திர மொழிபெயர்ப்பின் விரைவான முன்னேற்றத்தின் அடிப்படையில் தான் நிறைவேற்ற இயலும். இது நடைமுறையில் சாத்தியமில்லாத செயல்பாடாக தற்போது தோன்றினாலும் நாம் இதைச் சாத்தியமாக செய்கின்ற இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவதில ஆர்வம் காட்டணேடும். இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை ஒரு வடிவிலிருந்து மற்றொருவடிவுக்கு (அதாவது சொல்லிலிருந்து சொல்லையோ கூட்டுச் சொற்களிலிருந்து சுட்டுச் சொற்களையோ தொடரிலிருந்து தொடரையோ வாக்கியத்திலிருந்து வாக்கியத்தையோ) இடப்பெயர்ச்சிசெய்யும் அல்லது ஒரு மொழியிலுள்ள உரைகளை மற்றொரு மொழியில் உருப்படுத்தம் செய்யும் எளிய முறையோ அல்ல. இயந்திர மொழிபெயர்ப்பு உருவாக்கப்பட்ட உரை இலக்குமொழியில் இலக்கண அடிப்படையில் சரியான மற்றும் கருத்துரு அடிப்படையில் ஏற்றுக் கொள்ளத்தக்க வகையில் அமைய உறுதி தருவதைக் குறிப்பிடுகின்றது. இயந்திர மொழிபெயர்ப்பு அர்த்தம் மற்றும் பொருளடக்க அடிப்படையில் ஒன்றாக இல்லாவிடினும் குறைந்தது மூலமொழிக்கு அண்மையில் இருக்க வேண்டும். மூல மொழியிலுள்ள தகவல் இலக்கு மொழியில் மாற்றப்படும் போது இழப்பு எற்படக் கூடாது. மொழிப்பில் மூலமொழியை விடக் கூடுதல் தகவல் சேர்க்கப்படலாகாது. இயந்திர மொழிபெயர்ப்பில் துல்லியத்தைப் பெற இவைகள் தடைகளாகும். இயந்திர மொழிபெயர்ப்பு உருபனியல், சொல்லியல், தொடரியல் பொருண்மையியல், பயன்வழியியல், கருத்தாடல், புலனறிவு இவற்றைச் சார்ந்திருக்க வேண்டும். இதன் அர்த்தம் என்னவென்றால் ஒரு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை ஒரு மனித மொழிபெயர்ப்பாளருக்குள்ள திறனைப் பெற்றிருக்க வேண்டும் என்பதாகும். இதைச் சாதிப்பதற்கு நாம் காத்ததிருக்க வேண்டும். இருப்பினும் இரு மொழி அல்லது இணை மொழி பெயர்ப்பு விரிதரவுகள் இதைச் செய்வதற்குச் சாத்தியமான சூழலைத் தருகிறது. இது நம்மை விரிதரவுகள் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அணுகுமுறைக்குக் கொண்டு செல்கிறது. இதன் காரணமாக இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையின் திறன் கூடுகிறது. வெளியீடானது ஏற்றுக் கொள்ளத் தக்க முறையில் பின் திருத்தம் செய்யப்படுவதன் காரணமாகவும் குயைான தகவலின் மூலமாகத் திருத்தப்படாத வடிவில் பயன்படுத்தப்படுவதன் காலரணமாகவும் இயந்திர மொழிபெயர்ப்பு முற்றிலும் சரியான மொழி பெயர்ப்பை உருவாக்குவதை நோக்கமாகக் கொண்டதல்ல எனக் கூறவியலும். இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை மொழியியல் அடிப்படையிலும் புலனறிவு அடிப்படையிலும் உத்தேச ஒழுங்குமுறையைத்தான் நோக்கமாகக் கொள்கிறது. முயற்சித்தல் மற்றும் பிழை செய்தல் என்ற நீண்ட செயற்பாங்குகளின் வழி இலக்கை அடையலாம்.

8.2 வரலாற்றிலிருந்து கிடைக்கின்ற பாடம் (Lessons learnt from History)

1950-களில் இயந்திர மொழி பெயர்ப்பு உலகிலுள்ள பெரும்பாலான எல்லா மொழிகளுக்கும் நேரடியாகவோ மறைமுகமாகவோ மூல மொழியிலிருந்து இலக்கு மொழிக்கு இயந்திர மொழிபெயர்ப்பு செய்ய முயற்சிகள் எடுக்கப்பட்டன. இதுவரை முழு வெற்றி அடைய இயலவில்லை. இதன் அர்த்தம் என்னவென்றால் மூல மொழிக்கும் இலக்கு மொழிக்கும் மொழி பயன்பாட்டாளர்களால் ஏற்றுக் கொள்ளக்கூடிய மொழிபெயர்ப்புகளைத் தானியக்கமாக உருவாக்கும் ஒரு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை இல்லை. இதன் விளைவாக இயந்திர மொழிபெயர்ப்பு ஆய்வாளர்கள் மரபு சார்ந்த விதி அடிப்படையிலான அணுகுமுறைக்குத் தங்கள் கவனத்தை திருப்பியுள்ளனர்.

8.3 விரிதரவு அடிப்படையிலான அணுகுமுறை (The corpus-Based Approach)

கடந்த நூற்றாண்டுகளாக இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளுக்கு வேண்டி உருவாக்கப்பட்ட விதிகளின் குழுமம் மொழி அமைப்பை முழுவதுமாகப் பிரதிபலிக்கவில்லை. விதிகளின் சில குழுமங்களால் இயற்கை மொழியின் நெகிழ்வான அமைப்பை உருப்படுத்தம் செய்ய இயலாது. இதன் காரணமாக விரிதரவு அடிப்படையிலான இயந்திர மொழி பெயர்ப்பு உருவானது. இது பின்வரும் இயந்திர மொழிபெயர்ப்பு வகைகளை உள்ளடக்கும்.

1. எடுத்துக்காட்டு அடிப்படையிலானக இயந்திர மொழிபெயர்ப்பு
   (Example-based Machine Translation)
2. புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு
   (Statistics-based Machine Translation)

விரிதரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் விளைவுகளில் ஒன்று இருமொழி உரைகளின் வாக்கியங்களை வரிசைப்படுத்தும் வழிமுறை வரைவுகளின் உருவாக்கமாகும். இவ்வெளிய விளைவு இயந்திர உதவி வழி மனித மொழி பெயர்ப்பின் அடிப்படை விஷயங்களில் ஒன்றாக அமைந்தது; ஏனென்றால் இது மொழிபெயர்ப்புக்கு உதவும் பல புதிய கருவிகளுக்குப் பொருத்தமான அடிப்படையாக அமைந்தது. விரிதரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு மனித மொழிபெயர்ப்பாளர்களால் ஏற்கனவே உருவாக்கப்பட்ட மொழிபெயர்ப்புகளை பயன்படுத்தி முழுமையாகவோ பகுதியாகவோ அவற்றின் அக அமைப்புகளைக் கண்டுபிடிக்க முயலுகிறது. இப்பகுப்பாய்வு அடிப்படையிலான அணுகுமுறை மொழி பெயர்ப்பாளர்களுக்கு உதவும் கருவிகளை உருவாக்க உதவியது. இருமொழி அல்லது இணைமொழி பெயர்ப்பு விரிதரவை இயந்திர மொழிபெயர்ப்பில் பயன்படுத்தும் கருத்து புதியது

அல்ல. இயந்திர மொழிபெயர்பின் ஆரம்ப காலத்திலேயே இது முயற்சிக்கப்பட்டது. பொதுவாக இணை விரிதரவுகள் ஒரு மொழி விரிதரவுகளை விட மொழிகளைப் பற்றிய விரிவான தகவலைக் கொண்டிருக்கிறது. விரிதரவு ஒழுங்குமுறை எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அணுகுமுறையையும் புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அணுகுமுறையும் இணைத்து பயன்படுத்துவதால் இது உயர்ந்ததாக இருக்கிறது. விரிதரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு மூல மொழியிலும் இலக்கு மொழியிலும் உள்ள விரிதரவுகளில் முதன்மையான பகுப்பாய்விலிருந்து உருவாக்கப்பட்ட வகைகளின் பரப்பெல்லைகளின் அடித்தளத்தில் அமைந்துள்ளது. பகுப்பாய்வு மொழிபெயர்க்கப்பட்ட விரிதரவில் அடங்கியுள்ள தொடர்கள், மரபுத் தொடர்கள், வாக்கியங்கள், பத்திகள் மற்றும் பிற மொழியியல் பண்புக்கூறுகளின் உருபனியல், பொருண்மையியல் மற்றும் புலனறி பொருள்கோள்களை உட்படுத்தும்.

விரிதரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் அடிப்படை நெறிமுறை மொழி பெயர்ப்புக்கு முன் வரையறுக்கப்பட்ட தீர்மானங்கள் இல்லை. ஆனால், மனிதரால் மொழி பெயர்க்கப்பட்ட உரைகளில் மிக சாத்தியமான தீர்மானங்கள் காணப்படும் என்ற நம்பிக்கையின் அடிப்படையில் அமைவதாகும். அதாவது மனித மொழிபெயர்ப்பாளர்களின் திறமையின் பெரும்பகுதி மொழி பெயர்க்கப்பட்ட உரைகளில் காணப்படும் மொழி நிகரன்களில் குறியாக்கம் செய்யப்பட்டுள்ளன என்று கூறலாம். எல்லைகுட்பட்ட களங்களில் இவ்வணுகுமுறையால் பெறப்பட்ட சமீபகால வெற்றி பொதுவான களங்களிலும் இம்மாதிரியான வெற்றியைப் பெற மொழி சார்ந்த மற்றும் மொழிக்குப் புறம்பான அறிவுதேவை என்று காட்டுகின்றது. எல்லா களங்களிலும் இவ்வணுகுமுறையின் வெற்றி குறித்து தீர்மானமாக வருவதுரைக்க இயலாது. இருப்பினும் நன்றாகத் திட்டமிடப்பட்ட மொழிபெயர்ப்பு விரிதரவிலிருந்து பெறப்பட்ட தகவலிருந்து உருவாக்கப்பட்ட நெறிமுறை எல்லைக்குட்பட்ட மற்றும் பொது களங்களில் வெற்றியைத் தரும் என்று கூறவுயலும்.

8.4 விரிதரவு அடிப்படையிலான அணுகுமுறையுடன் தொடர்புடைய சிக்கல்கள்

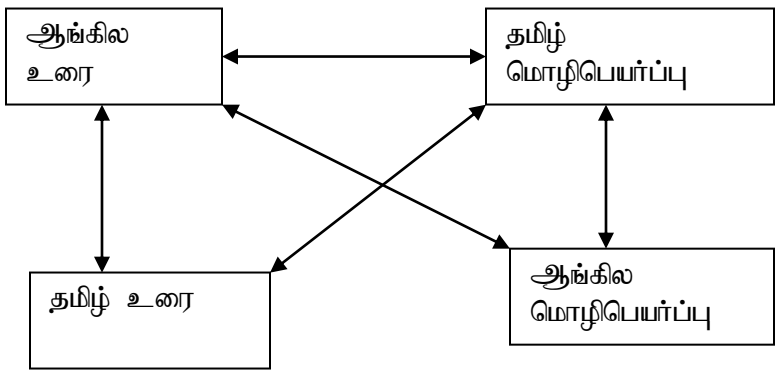
(Issues Related With Corpus Based Approach)

விரிதரவு அடிப்படையிலான மொழிபெயர்ப்பு விரிவான விரிதரவு பகுப்பாய்வின் மற்றும் ஆய்வின் வழி கூறத்தக்க முன்னேற்றத்தைப் பெற்றுள்ளது. இந்நாள் வரை பல்வேறு வகையிலான மொழிபெயர்ப்பு விரிதரவுகள் பொருத்தமான ஒழுங்குமுறைகளைத் திட்டமிட உதவும் முக்கியமான உள்ளறிவைத் தரவேண்டி உருவாக்கப்பட்டு பகுப்பாயப்பட்டுள்ளன. பொதுவாக இம்மொழிபெயர்ப்பு விரிதரவு இயற்கையாகப் பெறப்படும் மொழித் தரவுகளின் பெரிய சேகரிப்பை உருப்படுத்தம் செய்கின்றது.

இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் மொழிபெயர்ப்பு விரிதரவின் பயன்பாடு கட்டாயமானதாகும் என்று தெளிவாக்குகின்றது; ஏனென்றால் அவை மொழி மற்றும் மொழிக்குப் புறம்மான எடுத்துக்காட்டுகளின் மற்றும் தகவல்களின் பல்வேறு விதமான வகைகளை நமக்குத் தருகிறது.

8.4.1 மொழிபெயர்ப்பின் விரிதரவு உருவாக்கம் (Generation of Translation Corpora)

மொழிபெயர்ப்பு விரிதரவு மூல மொழியின் உண்மையான பண்புக் கூறுகளையும் இலக்கு மொழியிலிருந்து பெறப்பட்ட அவற்றின் மொழி பெயர்ப்புகளையும் கொண்டிருக்கும். இவ்விரிதரவு மொழிகளைக் கடந்து சொற்களின் மற்றும் தொடர்களின் பொருண்மையையும் செயல்பாட்டையும் கொண்டிருக்கும். இதன் காரணமாக ஒரே கட்டுபாட்டிற்குள் இரு வேறுபட்ட மொழிகளின் குறிப்பிட்ட பொருண்மைகளின் மெய்யுருவாக்கத்தை ஒப்பிடுவதற்குப் பொருத்தமான அடிப்படையைத் தருகின்றது. மேலும் அவை மொழிகளைக் கடந்த வேறுபட்ட சொற்களைக் கண்டுபிடிப்பதைச் சாத்தியமாக்குகிறது. இவ்வாறு மொழிபெயர்ப்பு விரிதரவு மொழிகளைக் கடந்த தரவுப் பகுப்பாய்வுக்கும் மொழிபெயர்ப்புக்குத் தேவையான விதி வடிவாக்கத்திற்கும் பயனுள்ள கூடுதல் மூலவளங்களைத் தருகின்றது. மொழிபெயர்ப்பு விரிதரவின் உருவாக்கம் ஒரு சிக்கலான வேலையாகும். இது விரிதரவின் உருவாக்கம் மற்றும் பகுப்பாய்வி இவற்றின் நீண்ட அனுபவம் உள்ள விரிதரவு மொழியிலாரின் நிரந்தரமான மற்றும் கவனமான வழிகாட்டலை வேண்டுகிறது. மொழிபெயர்ப்பு விரிதரவு ஒப்பீட்டு மற்றும் இணை விரிதரவுகளின் (Comparable Parallel Corpora) முன்னேற்றத்திற்கான வழிகளை ஒன்றிணைக்க வேண்டி தகுதியுடையதாக ஆக்க உருவாக்கப்படும். இரு மொழிகளிலிருந்து உரை மாதிரிகள், உரை வகை, பாடப்பொருள், நோக்கம் மற்றும் கலைச்சொல் இவற்றின் அடிப்படையில் இயன்ற வரைக்கும் பொருத்தப்படும். இரு மொழிகளுக்கிடையிலுள்ள மொழிபெயர்ப்பு விரிதரவின் அமைப்பு பின்வருமாறு அமையும்:



இம்மாதிரியான விரிதரவுக்கு உரை மாதிரிகள் தெரிவு பின்வரும் கொள்கைகளால் பொதுவாக வழி நடத்தப்படும்:

1. எழுதப்பட்ட உரைகளில் பயன்படுத்தப்படும் மொழி மாதிரிகள் மட்டும் தான் மொழி பெயர்ப்பு விரிதரவில் உட்படுத்தப்பட்டுள்ளன.
2. பேச்சு கருத்தாடலிலிருந்து பெறப்படும் உரைகளை உட்படுத்தும் சந்தர்ப்பம் இல்லை. ஏனென்றால் தற்போதைய இயந்திர மொழிபெயர்ப்பு எழுதப்பட்ட உரைகளை மட்டுமே இலக்காகக் கொண்டுள்ளது. உட்படுத்தப்பட்டள்ள உரைகள் தற்கால மொழியைப் பிரதிபலிப்பது எதிர்பார்க்கப்படும். இருப்பினும், பழங்கால நூல்கள் வரலாற்று உரைகளின் மொழி பெயர்ப்புக்குத் தேவையானதாகும்.
3. மொழிபெயர்ப்பு விரிதரவு எந்தக் குறிப்பிட்ட வட்டார மொழிக்கோ மொழி வகைக்கோ உரிய உரை வகைக்கு எல்லைப்படுத்தப்படவில்லை. அவை மொழி பயன்பாட்டின் எல்லாச் சாத்தியமான களங்கள் மற்றும் துறைகளிலிருந்து பெறப்பட்ட உரை வகைளின் பரந்த பரப்பெல்லையை உட்படுத்தும்.
4. இரண்டு மொழிகளிலிருந்து எடுக்கப்பட்ட உரைகள் இயன்றவரை ஒப்பிட தக்கவையாக இருக்கும். அவை இனம் (எ.கா. செய்தி), வகை (எ.கா.அரசியல்), பொருளடக்கம் (எ.கா. தேர்தல்) மற்றும் வடிவம் (எ.கா. அறிக்கை) இவற்றின் அடிப்படையில் பொருத்தம் பார்க்கப்படும். அவைகள் நோக்கம் பயன்படுத்துபவரின் மொழி இவற்றின் அடிப்படையிலும் பொருத்தமுடையதாக இருக்கும்.
5. மொழிபெயர்ப்பு விரிதரவில் உட்படுத்தப்பட்டுள்ள மாதிரிகள் ஒரு இயல்பான தடை நிலையில் (எ.கா. பகுதி, துணைப்பகுதி, அல்லது பிரிவு, பத்தி போன்றவை) தொடக்கத்திலிருந்து இறுதிவரை எடுக்கப்பட்ட உரைகளில் விரிவான மற்றும் இயற்கையான பாகங்களைக் கொண்டிருக்கும்.

8.4.2 மொழிபெயர்ப்பு விரிதரவைப் பொருத்தமாக வரிசைப்படுத்துதல்

மொழிபெயர்ப்பு விரிதரவுகளைப் பொருத்தமாக வரிசைப்படுத்துவது என்பது மூலமொழியின் ஒவ்வொரு மொழிபெயர்ப்பு அலகும், இலக்குமொழியின் ஒவ்வொரு மொழிபெயர்ப்பு அலகுடன் பொருந்த வேண்டும் என்பதாகும். இங்கு மொழிபெயர்ப்பு அலகு என்பது சிறிய தொடர்ச்சிகளான சொற்கள்,

கூட்டுச்சொற்கள், தொடர்கள் மற்றும் வாக்கியங்கள் என்பனவற்றுடன் பெரிய தொடர்ச்சிகளான பத்திகள், பகுதிகள் என்பனவற்றையும் உள்ளடக்கும். மொழிபெயர்ப்பு அலகுகளின் தெரிவு மொழியியல் ஆய்வுக்குத் தெரிந்தெடுக்கப்பட்டுள்ள பார்வையையும், பயன்படுத்தப்பட்டுள்ள விரிதரவின் வகையையும் பொறுத்து அமையும். மொழி பெயர்க்கப்பட்ட விரிதரவு மூலத்திலிருந்து (எ.கா. சட்டம் மற்றும் தொழில்நுட்ப விரிதரவுகள்) மாறாத உயர்ந்த நிலையை வேண்டினால் இரு விரிதரவுகளின் நெருங்கிய வரிசைப் பொருத்தம் வாக்கியங்கள், சொற்கள் என்ற அளவில் விலகி நிற்கும். விரிதரவு தழுவலாக இருந்தால் மூலத்தின் நேரடி மொழிபெயர்ப்பாக அமையாமல் பத்திகள், பகுதிகள் என்ற பெரிய அலகுகளை வரிசைப்பொருத்தம் செய்ய முயற்சிக்கப்படும். இவ்வாறு வரிசைப்படுத்தத்தின் செயற்பாடு பயன்படுத்தப்பட்டுள்ள விரிதரவின் வகை அடிப்படையில் நேர்படுத்தப்படும். வரிசையாக வருவதும் மனித மொழிபெயர்ப்பாளர்களின் நம்பகத் தன்மையும் மொழிபெயர்க்கப்பட்ட விரிதரவுகளை வரிசைப்பொருத்தம் செய்ய உதவும். இது கூடுதல் தொழில்நுட்ப விரிதரவுக்குப் பகுதி உண்மையாகும். மாறாக, இலக்கிய விரிதரவுகள், விரிதரவில் பயன்படுத்தப்படும் நிகரன்கள் முன்னரே முறையாக்கம் செய்யப்பட்டிருந்தால் வாக்கிய மட்டத்திற்கு கீழே உள்ள அலகுகளின் வரிசைப் பொருத்தத்திற்கு உட்படும். எந்த நிலையிலும் (பத்தி, வாக்கியம், சொல் போன்றவை) மொழிபெயர்க்கப்பட்ட விரிதரவுகள் இணையான அலகுகள் கொண்ட எளிய சொல் தரவு மையங்களாகக் கருதப்படுகிறது. இதன் முக்கிய நோக்கம் இரண்டு மொழிகளுக்கிடையே காணப்படும் அமைப்பு ஒற்றுமைகளைக் காட்டுவதற்கு அல்ல; பயன்வழியியல் அடிப்படையில் மூல உரை அலகுகளுக்கு அண்மைப்பட்டதாகத் தோன்றும் இலக்கு உரை அலகுகளைத் தேடுவதாகும். இதைச் செய்வதற்குத் தொடக்க நிலை, இருமொழிய அகராதிகளின் உதவியால் சொற்களின் ஆரம்ப வரிசைப் பொருத்தத்தைச் செய்வதாகும். இம்மாதிரியான குறைபாடான வரிசைப் பொருத்தங்கள் வாக்கிய நிலைகளில் திருப்திகரமான விளைவுகளைத் தரும்.

8.4.3 இந்திய மொழிகளின் இன்றைய நிலை

விரிதரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பில் முக்கியமான தடைகளில் ஒன்று இருமொழிய மொழிபெயர்ப்பு விரிதரவுகளின் உருவாக்கமாகும். ஒருமொழிய விரிதரவில் போதுமான அளவு எழுதப்பட்ட உரைகள் இருந்தாலும் பின்வரும் செயல்படுத்துநிலை சிக்கல்களின் காரணமாக இருமொழிய மொழிபெயர்ப்பு விரிதரவுகளை மறுதிட்டவரைவு செய்வதற்கு ஆர்வமுள்ள முயற்சிகள் எடுக்கப்படவில்லை:

உருவாக்கம்: இந்திய மொழிகளுக்கிடையே வேறுபட்ட வகையான மொழிபெயர்ப்பு விரிதரவு நீண்டநாள் தேவையாகும். சில ஒருமொழிய விரிதரவுகள் ஒரேமாதிரியான திட்டவரைவுக் கொள்கைகளையும் உரை வகைகளையும் கொண்டு உருவாக்கப்பட்டுள்ளன. இவற்றை முறையாக மொழிபெயர்ப்பு விரிதரவுகளை உருவாக்கப் பயன்படுத்தலாம்.

ஏற்றுக்கொள்ளல்: மனித முயற்சியால் உருவாக்கப்பட்ட மொழிபெயர்ப்பு விரிதரவுகள் சில (எ.கா. National Book Trust மற்றும் Sahitya Academy போன்றவற்றால் வெளியிடப்பட்ட பல உரைகள்) ஏற்கனவே உருவாக்கப்பட்டுள்ளது என்றாலும் அவை ஏற்றுக்கொள்ளக்கூடிய மின்வடிவில் இல்லை.

மாற்றம்: வேறுபட்ட மூலம், உரைவகை, மாதிரி மற்றும் வடிவமைப்பு கொண்ட பிற விரிதரவுகள் இணை மொழிபெயர்ப்பு விரிதரவுகளாக மாற்ற வேண்டி ஆயப்படவேண்டும்.

சுத்தமாக்கல்: மொழிபெயர்ப்பு விரிதரவு எதிர்கால செயற்பாங்கிற்கும் ஏற்றுக்கொள்ளலுக்கும் பயன்பாட்டிற்கும் வேண்டி இயந்திரப் பயன்பாட்டு வடிவமைப்பில் வைத்திருக்க மொழி வல்லுனர்களால் முறையாக தலையிடப்படவும் கட்டுப்படுத்தப்படவும் வேண்டும்.

இணைப்பொருத்தமாக வரிசைப்படுத்தல்: மொழிகளுக்கிடையில் பொருந்தும் பகுதிகளை அடையாளங்காண வேண்டி மொழிபெயர்ப்பு விரிதரவு இணைப்பொருத்தமாக வரிசைப்படத்தப்படவேண்டும்.

8.5. மொழிபெயர்ப்பு விரிதரவுகளில் மொழியியல் செயல்பாடுகள்
(Linguistics Tasks On Translation Corpora)

மொழிபெயர்ப்பு விரிதரவு தொகுத்தலுக்கும் வரிசைப்படுத்தலுக்கம் பின்னர் மொழியியல் பகுப்பாய்வின் வேறுபட்ட நிலைகளுக்கு உள்ளாக்கப்படுவதற்கு முன்னர் வரிசைப் பொருத்தத்தின் பல நிலைகளைக் கடக்கும். இது மொழி பெயர்ப்பு நிகரன்களின் முறையாக்கத்தின் (வடிவாக்கத்தின்) நிலைகளுக்கு அடிப்படையாக அமையும். பொதுவாக மொழியியல் ஆய்வு பின்வருவனவற்றை உள்ளடக்கும்:

1. உருபன்களின் வடிவத்தையும் செயல்பாட்டையும் அடையாளம் காண உதவும் சொற்களின் உருபனியல் ஆய்வு
2. தொடர்புள்ள விரிதரவுகளில் தொடர்களின் வடிவத்தையும் செயல்பாடுகளையும் அடையாளம் காண உதவும் தொடரியல் ஆய்வு

3. விரிதரவுகளில் பயன்படுத்தப்பட்டுள்ள சொற்களின் புறவடிவங்களுக்கு இடையில் உள்ள இடைமுகத்தை ஆய உதவும் உருபனியல்-தொடரியல் ஆய்வு. இதன் துல்லியமான மற்றும் திறமையான ஆய்வு பகுப்பாய்வின் பண்பையும், விரைவையும் அதிகரிக்கும்.
4. சொற்கள், தொடர்கள் போன்ற அலகுகளின் பொருண்மையைக் கண்டுபிடிக்கவும் இதில் வரும் பொருள் மயக்கங்களைக் கண்டுபிடிக்கவும் உதவும் பொருண்மையில் ஆய்வு

8.5.1 மொழிபெயர்ப்பு ஆய்வு (Translation Analysis)

இயந்திர மொழிபெயர்ப்பு ஆய்வாளருக்கு இடையிலான முக்கியமான வாக்குவாதம் மொழிபெயர்ப்பு விரிதரவுகளின் ஆய்வில் உட்படும் கலவைத் தன்மையின் நிலையைப் பற்றியாகும். பொதுவான நம்பிக்கை இயற்கையான உரைகளில் வரும் பல எண்ணிக்கையிலான மொழியியல் நடத்தைகள் ஆயப்பட்டு வெளிப்படையாக உருப்படுத்தம் செய்யப்படும் வரை உயர்ந்த தன்மையுள்ள இயந்திர மொழிபெயர்ப்பு எய்த இயலாததாகும் என்பதாகும். சொல் மயக்கம் மற்றும் உறுப்பு பொருத்தம் இவற்றின் சிக்கல்களை மொழிபெயர்ப்பு விரிதரவுகளிலிருந்து பெறப்பட்டு உட்படுத்தப்பட்ட ஒரு மொழியின் அகராதியலும் இலக்கணத்திலும் சேகரித்து வைக்கப்பட்டுள்ள ஏரளமான அறிவின் உதவியால் தீர்க்க இயலும். இது மொழிபெயர்ப்பு ஆய்வின் கடுமையான செயற்பாங்கின் பயன்படுத்தலை வேண்டும். சுருக்கமாக இயந்திர மொழிபெயர்ப்பில் மொழி ஆய்வு நுட்பத்திற்குப் பல பயன்பாடுகள் இருக்கின்றன. அவை பின்வருமாறு அமையும்:

1. மொழிபெயர்ப்பு பகுப்பாய்வு நுட்பம் முன்னரே இருக்கக் கூடிய மொழி பெயர்ப்புகளை அமைப்பாக்கம் செய்ய உதவும். இதன் மூலம் அவை புதிய மொழி பெயர்ப்புகளின் உருவாக்கத்தில் மறு உபயோகம் செய்ய இயலும்.
2. மொழிபெயர்ப்பு பகுப்பாய்வு தொழில் நுட்பம் மொழிபெயர்ப்பு தவறுகளின் சில வகைகளைக் கண்டுபிடிக்க வேண்டி தொடக்க நிலை மொழிபெயர்ப்புகளில் பயன்படுத்தப்படுகிறது.
3. மொழிபெயர்ப்பு பகுப்பாய்வு நுட்பம் இணை மொழிச் சொற்களால் ஏற்படும் குறுக்கீட்டு தவறுகளிலிருந்து சுதந்திரமாக மொழிபெயர்ப்பு இருந்தால் அதை பரிசோதிக்கப பயன்படுத்தப்படும்.

8.5.2 இருமொழிய அகராதியின் உருவாக்கம் (Building Bilingual Dictionary)

விரிதரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் மற்றொரு பயனுள்ள மற்றும் முக்கிய பகுதி இருமொழி அகராதிகளின் உருவாக்கமாகும். இதன் குறைபாடு இந்திய மொழிகளின் இயந்திர மொழிபெயர்ப்பின் தடைகளில் ஒன்றாகும். மரபு அகராதிகள் இக்குறைபாட்டை நிவர்த்தி செய்ய இயலாது; ஏனென்றால் அவை சொல் துணை வகைப்பாடு, சொல்தெரிவு, சொல்தேர்வு கட்டுப்பாடு மற்றும் சொல் அலகுகளின் பயன்பாட்டு களங்கள் இவற்றைப் பற்றிய போதுமானத் தகவல்களை கொண்டிருக்கவில்லை. இருமொழிய அகராதி உருவாக்கத்திற்கு நாம் பல வழிகளில் அடையாளப்படுத்தப்பட்ட மற்றும் பகுத்துக் குறிக்கப்பட்ட விரிதரவுகளில் பயன்படுத்தப்பட்டுள்ள பல்வேறு புள்ளியியல் நெறிமுறைகளைப் பயன்படுத்த வேண்டும்.

8.5.3 மொழிபெயர்ப்பு நிகரன்களின் பிரித்தெடுப்பு (Extraction of Translation Equivalents)

மொழிபெயர்ப்பு விரிதரவுகளில் மொழிபெயர்ப்பு நிகரன்கள் தேடல் மூல மொழியில் ஒரு குறிப்பிட்ட பொருண்மையை அல்லது கருத்துருவை வெளிப்படுத்தும் ஒரு குறிப்பிட்ட வடிவிலிருந்து தொடங்கப்பட வேண்டும். மூல மொழியில் இச்சொல் கண்டுபிடிக்கப்பட்டு தனியான தரவு மையத்தில் பதிவு செய்யப்பட்ட பின் தேடல் செயற்பாங்கின் 2-வது பகுதி தொடங்கும். இது இலக்குமொழியில் ஒரே அர்த்தத்தைக் கொண்டிருக்கிற சொல்லின் அடையாளம் காணலை உட்படுத்தும். இயல்பாகப் பெரும்பாலான மொழிபெயர்ப்பு விரிதரவுகள் மொழிபெயர்ப்பு நிகரன்களின் மிகுந்த பரப்பெல்லையை வெளிப்படுத்த இயலும். அவை மாற்று வடிவின் பரிசோதனைக்கு முக்கியமான மூலமாக அமையும். இருப்பினும் பொருத்தமான நிகரன்களின் தேர்வை உறுதி செய்யும் காரணிகள், சொற்களின் பயன்பாட்டின் மறு நிகழ்வு அமைப்பொழுங்கின் அடிப்படையில் உறுதி செய்யப்படும்.

8.5.4 கலைச்சொல் தகவல் வங்கியின் உருவாக்கம் (Generation Of Terminology Databank)

விரிதரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்புக்குத் தொழில் நுட்ப மற்றும் அறிவியல் கலைச்சொற்களில் பொருத்தமான ஆக்கத்திற்குக் கலைச்சொல் ஆக்கவியலாரின் எண்ணத்தின் அடிப்படையிலான ஆய்வுத் தேவை. கலைச்சொல் ஆக்கவியலாரின் அடிப்படைச் செயல்பாடு இலக்கு மொழிக்கு அயல் கருத்துக்களையும் கருத்துருக்களையும் உருப்படுத்தம் செய்ய மிகப் பொருத்தமான அல்லது ஓரளவுக்குப் பொருத்தமான சொற்களைத் தேர்வு செய்வதாகும். இதைச் செய்வதற்குக் கலைச்சொல் ஆக்கவியலார் இலக்குமொழி பயன்படுத்துபவர்களுக்கிடையில் கலைச்சொற்களின் பொருத்தமான தன்மை, இலக்கணத் தன்மை, ஏற்றுக்கொள்கை மற்றும் பயன்பாடு போன்ற பல்வேறு

காரணிகளை மனதில் கொள்ள வேண்டும். முக்கியமான காரணி கலைச்சொற்களின் தரம் ஆகும். இதன்படி மொழியியல் கொள்கைகளின் அடிப்படையில் பொருத்தமான வழிகளைப் பயன்படுத்தி புதிய சொற்கள் உருவாக்கப்படும். மொழிபெயர்ப்பு விரிதரவு ஒரு கருத்து, நிகழ்ச்சி, விஷயம் அல்லது கருத்துரு என்பதை உருப்படுத்தம் செய்யப் பலரால் உருவாக்கப்பட்ட பல கலைச்சொற்களின் பெரிய பட்டியலிலிருந்து பொருத்தமான ஒன்றைத் தேர்வு செய்யும் முக்கிய பங்களிப்பைச் செய்கின்றது. புதிய சொற்கள் உருவாக்கத்தின் தொடர்ச்சியான முயற்சி விரிதரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அமைப்பைத் திட்டமிடுபவர்களுக்கு ஒன்றுக்குப் பதிலாக ஒன்றைத் தேர்ந்தெடுப்பதில் சிக்கலைத் தரும். புதியச் சொல்லை உருவாக்குவதா அல்லது இலக்குமொழியில் பயன்பாட்டால் இயல்பாக்கம் செய்யப்பட்ட மூலமொழிச் சொல்லைப் பயன்படுத்துவதா என்ற விவாதம் எழும். சில தொழில நுட்பச் சொற்கள் அவற்றின் உண்மையான மூலத்தைக் கண்டுபிடிக்க இயலாதவாறு இயல்பாக்கம் செய்யப்பட்டிருக்கும். அன்றாட வாழ்க்கையில் அவற்றைப் பயன்படுத்தும் உயர்ந்த நிகழ்வெண் அவற்றைச் சொற்றொகையின் பகுதியாக மாற்றும். எனவே மொழிபெயர்பின் போது அவற்றை இடம் பெயர்க்கத் தேவையில்லை. பொதுவாக இலக்கு மொழியில் மொழி பெயர்க்கப்பட்ட விரிதரவு மூல மொழியிலிருந்து கடன் வாங்கப்பட்ட புதிய கருத்துக்களையும் கருத்துருக்களையும் வெளிப்படுத்த பொருத்தமானச் சொற்களைத் தேர்ந்தெடுக்க நல்ல மூலவளமாகும்.

8.5.5 சொல் தேர்வு கட்டுப்பாடு (Lexical Selection Restriction)

மூலமொழியிலிருந்து தெரிந்தெடுக்கப்பட்ட மிகப் பொருத்தமான நிகரன்களாகக் கருதப்பட வேண்டிய இலக்கு மொழிச் சொற்களின் தெரிவு மூலமொழியிலும் இலக்குமொழியிலும் பட்டறிவுள்ள மொழியியலாரின் திறமையான மற்றும் கவனமான பொருள்கோளை வேண்டும் மற்றொரு கலவைத்தன்மையான செயல்பாடாகும். இதன் பொருள் என்னவென்றால் மொழியியலார் இலக்கு மொழியல் இருக்கின்ற ஒரே பொருண்மையை தருகின்ற வடிவங்களின் பெரிய தொகுப்பிலிருந்து பொருத்தமான சொல்லைத் தேர்ந்தெடுக்க வேண்டும் என்பதாகும். பின்வரும் எடுத்துக்காட்டுகளைக் கவனத்தில் கொள்ளவும்:

ஆங்கிலம் : That saint ate food

தமிழ் : அந்த முனிவர் போஜனம் செய்தார்.

ஆங்கிலம் : He ate food

தமிழ் : அவன் உணவு சாப்பிட்டான்.

மேற்சொன்ன எடுத்துக்காட்டுகளில் நாம் தமிழில் எழுவரின் பண்பின் அடிப்படையில் சரியான இணைச் சொல்லைத் தெரிந்தெடுக்க வேண்டியுள்ளது. மொழியியலாரின் முக்கியமான செயல்பாடு இரு மொழிகளுக்கு இடையிலான மொழிபெயர்ப்பில் வேறுபட்ட சமூக மொழியியல் காரணிகளைப் பொருத்தமான சொல் அலகுகளைக் காண்பதாகும். இம்மாதிரியான எடுத்துக்காட்டுகள் வெற்றிகரமான மற்றும் அறிவு பூர்வமான மொழிபெயர்ப்பில் சொல் தேர்வு கவனமாகக் கையாளப்பட வேண்டும் என்பதைக் காட்டப்படுகிறது. இச்சிக்கல் மனித மொழிபெயர்ப்பில் கவனமாகக் கையாளப்பட்டாலும் பெரும்பாலான இயந்திர மொழிபெயர்ப்பில் புறக்கணிக்கப்படுகின்றன. இச்சிக்கலை தீர்பதற்கான நல்ல வழி இயந்திரத்தால் படிக்க இயலும் அகராதியின் தனியான அறையில் பொருண்மையில் ஒற்றுமையைக் காட்டும் எல்லா வடிவங்களைப் பட்டியலிடுவதாகும். இம்மாதிரியான தரவுமையங்களை மொழிபெயர்ப்பு விரிதரவுகளிலிருந்து பிரிந்தெடுக்க இயலும். பொதுவாக இயந்திரம் படிக்க இயலும் அகராதிகள் பல பாடப்பொருள் களங்களாகப் பிரிக்கப்பட்டிருக்கும். இதனால் ஒரு குறிப்பிட்ட பாடப்பொருளைச் சார்ந்த சொற்களைப் பற்றிய பொருத்தமான தகவலைப் பெற இயலும்.

8.5.6 சொல் மயக்கத்தை நீக்குதல் (Dissolving Lexical Ambiguity)

இயல்பான சூழலில் மனித கருத்துப் பரிமாற்ற ஒழுங்கமைப்பு மொழியைக் கருவியாக கொண்டு பேசுபவரிடமிருந்து கேட்பவருக்குச் செய்தியைப் பரிமாற்றம் செய்வதை உள்ளடக்கும். சில நேரங்களில் செய்தியின் பரிமாற்றம் பொருண்மை மயக்கத்திலிருந்து விடுபடாமல் இருக்கும். இப்பொருண்மை மயக்கம் மொழியில் மிகப் பரவலான நடப்பாகும். மயக்கமான சொல்லுடன் தொடர்புடைய அகப்பொருண்மையின் போதாமையால் அல்லது கருத்துப் பரிமாற்றத்தின் ஒரு குறிப்பிட்ட நிகழ்வில் கூற்றின் அமைப்பால் ஏற்படும். இவ்வாறு மயக்கங்கள் சொல் மயக்கம் (Lexical ambiguity) என்றும் குறிப்புரை மயக்கம் (Referential ambiguity) என்றும் தொடரியல் மயக்கம் (Syntactic ambiguity) என்றும் முன்றாகப் பகுக்கப்படும்:

They went to bank – சொல் மயக்கம்

He loves his wife – குறிப்புரை மயக்கம்

He saw a girl in the park with a binocular – தொடரியல் மயக்கம்

இயந்திர மொழிபெயர்ப்பு ஒழுங்கமைப்பு பேசுபவரின் மன உருப்படுத்தத்தின் புலனுணர்வு அடிப்படையில் உருவாக்கப்படுவதால் அது பேசுபவரால் பயன்படுத்தப்படும் சொற்களுக்கும் வாக்கியங்களுக்கும் எல்லைப்படுத்தப்பட்டிருக்கும். இதை நேர்செய்ய மொழிபெயர்ப்பாளர்கள் முல

அகராதியை ஒரு குறிப்பிட்ட சூழலில் பொருத்தமாகத் தோன்றுகின்ற நிகரனை இலக்கு அகராதியால் பொருத்துகிறார்கள். சில நேர்வுகளில் இலக்கு மொழியானது மூலமொழியின் சொல்லை உருப்படுத்தம் செய்ய இயலும் நிகரான சொல் அலகைக் கொண்டிருக்காது. இந்நேர்வுகளில் மொழிபெயர்ப்பாளர்கள் சொற்றொடரைப் பயன்படுத்த முயல்வர் அல்லது நிகரான விளக்கத்தைத் தர முயல்வர்.

விரிதரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பை ஆதரிப்பவர்கள் சொல் மயக்கச் சிக்கலை நீக்க இயற்கைமொழி உரைகளில் கூடுதலாக நிகழும் மயக்க வடிவுகளில் பெரும் எண்ணிக்கையை ஆய்ந்தும் வெளிப்படையாக உருப்படுத்தம் செய்தும் உயர்ந்த தரமான மொழிபெயர்ப்பைத் தரலாம் என்று வாதியிடுகின்றனர். சாத்தியம் என்றால் மயக்கமான வடிவுகளும் மொழிபெயர்ப்பு விரிதரவுகளிலிருந்து பெறப்பட்ட விரிந்த அறிவின் அடிப்படையில் ஆயப்பட்டு இயந்திரத்தால் புரிந்து கொள்ளப்படும் வடிவுகளும் அகராதியில் சேமித்து வைக்கப்பட வேண்டும். ஆனால் பகுத்தறிவாதிகள் இவ்விதம் மொழியியல் தகவல்களைப் பெறும் செயல்பாடு சரியானதும் அல்ல சாத்தியமானதும் அல்ல என்று வாதியிடுகின்றனர். களக்கட்டுப்பாடு மொழிபெயர்ப்பு (Domain–Specific Translation) மொழிபெயர்ப்புக்கு எப்போதும் ஆழமான பொருண்மையியல் ஆய்வு தேவையில்லை என்று வெளிப்படுத்துகின்றன. குறிப்பிட்ட மயக்கம் உள்ள சொல் அலகுகளின் ஆழமான பொருண்மையியல் ஆய்வுக்குப் பதிலாக நேரடியாக மூல மொழியிலிருந்து இலக்கு மொழிக்கு பதிலீடு செய்யும் அணுகுமுறையான எளிய ஆய்வைத் தெரிந்தெடுப்பது நல்லது. சில சூழல்களில் மயக்கங்கள் மொழி பெயர்ப்பின் போது அப்படியே கொண்டு செல்லப்படும் என்ற எதிர்பார்ப்பு அடிப்படையல் சில சொல் மயக்கங்களைப் புறக்கணித்தல் சாத்தியமானதும் தேவையானதுமாகும். இருப்பினும் சொல் மயக்கங்களின் ஆய்வு, இலக்குமொழியில் மயக்கமற்ற உருப்படுத்தங்களை உருவாக்கும் என்ற நிலையில் பொதுவான மொழிபெயர்ப்பு நேர்வுகளில் மயக்கத்தை புறக்கணிக்கக் கூடாது.

8.5.7 இலக்கண பொருத்தம் (Grammartical Mapping)

மொழிபெயர்ப்பில் இலக்கண பொருத்தம் என்பது மூல மொழியில் ஒரு சொல் இலக்கு மொழியில் உள்ள சொல்லுடன் அர்த்தம் உள்ள மொழிபெயர்ப்பைப் பெற வேண்டி பொருத்தப்படும். மொழிபெயர்ப்பில் பொருத்தத்திற்கு வேறுபட்ட திருத்தங்கள் இருக்கின்றன. அவை சொல் நிலைப் பொருத்தம், உருபன் நிலைப் பொருத்தம், இலக்கண நிலைப் பொருத்தம், தொடர்நிலைப் பொருத்தம், வாக்கியநிலைப் பொருத்தம் என்பனவாகும். மிகப் பொதுவானது மொழிபெயர்ப்புக்கு எடுத்துக்கொள்ளப்பட்ட இரு மொழிகளுக்கிடையில் வினை பங்களிப்பைப் பொருத்துவதாகும். இலக்கண பொருத்தம், வகைப்பாட்டியியல் அடிப்படையில் வேறுபட்ட இரண்டு மொழிகளுக்கிடையில் செய்யப்படும்

இயந்திர மொழி பெயர்ப்பில் முக்கியத்துவம் உள்ளதாகும். தற்போதைய சூழலில் நாம் ஆங்கிலத்திலிருந்து தமிழுக்கு மொழிபெயர்ப்பைப் பற்றி பேசும் பொழுது இது முக்கியத்துவம் வாய்ந்தது; என்னவென்றால் ஆங்கில0000000000000000000த்திற்கு SVO அமைப்பு உண்டு (He is Rama) ; தமிழுக்கு SOV அமைப்பு இருக்கிறது (அவன் சோறு சாப்பிட்டான்). எனவே இலக்கண பொருத்தமும் சொல் வரிசைமுறை மாற்றமும் இலக்கு மொழியில் வெளியீடுகளை வெளியிடுவதற்கு முக்கியத்துவம் வாய்ந்ததாகும்.

8.6. இறுதியுரை

இவ்வாறு விரிதரவு இயந்திர மொழிபெயர்ப்புக்குப் பெரிதும் பயனுள்ளதாக அமைகிறது. இணையான மொழிபெயர்க்கப்பட்ட விரிதரவு சொல் பொருத்தம், உருபன் பொருத்தம், இலக்கணப் பொருத்தம், தொடர் பொருத்தம், வாக்கியப் பொருத்தம் என்ற செயல்பாடுகளிலிருந்து மொழியின் சொல் மட்டம், உருபன் மட்டம், தொடரியல் மட்டம் என்ற நிலைகளில் நிகரன்களைக் கண்டறிந்து கூடுதலாக விரிதரவுக்குப் பயிற்சி தந்து இயந்திர மொழிபெயர்ப்பை மேம்படுத்தலாம்.

பகுதி 9

முடிவுரை

முந்தைய பகுதிகளில் விரிதரவின் விளக்கம், விரிதரவின் தொடக்க நிலை, விரிதரவின் இன்றைய நிலை, இதுவரை உருவாக்கப்பட்ட விரிதரவுகள், விரிதரவின் வகைப்பாடு, எழுத்துவடிவ விரிதரவின் உருவாக்கம், உரை விரிதரவின் ஆய்வு, மொழித் தொழில் நுட்பத்தில் விரிதரவு, முக்கிய மொழியியலில் விரிதரவு, இயந்திர மொழிபெயர்ப்பில் விரிதரவு, விரிதரவின் பயன்பாடுகள் என்ற தலைப்புகளில் விரிதரவு தொடர்பான செய்திகள் எல்லாம் விளக்கமாக எடுத்துரைக்கப்பட்டுள்ளன.

களமொழியலிருந்து விரிதரவு மொழியலுக்கான பயணம் பல மயில்கற்களைத் தாண்டி வந்துள்ளது. அன்றைய கால கட்டத்தில் மொழியியலார் கள ஆய்வை மேற்கொண்டு மொழித் தரவுகளைத் திரட்டி ஒரு மொழியின் அமைப்பை அறிய ஒலியியல், ஒலியனியல், உருபனியல், தொடரியல் மற்றும் பொருண்மையியல் மட்டங்களில் ஆய்வு செய்வார்கள். இத்தகைய ஆய்வு தொடக்க காலத்தில் அமைப்புமுறை மொழியியல் ஆய்வாக அமைந்தது. இது நடத்தையிலாரால் ஊக்கப்படுத்தப்பட்ட ஊன்றிக்கவனித்தல் அடிப்படையிலான அனுபவவாத அடிப்படையில் நிகழ்த்தப்பட்டது. சாம்ஸ்கி இத்தகைய ஆய்வுகளைக் செயல்திறன் சார்ந்தது என்றும் குறையுள்ளது என்றும் கூறிப் பகுத்தறிவு வாதத்திற்கு வித்திட்டார். இதன்பயனாக கள ஆய்வுகள் ஊக்கமிழந்து உள்ளுணர்வு ஆய்வுகள் ஊக்கமடைந்தன. சாம்ஸ்கி கள ஆய்வாளர்களால் திரட்டப்பட்ட தரவு குறையுடையது என்று வாதிட்டார். ஆனால் விரிதரவின் வரவால் சாம்ஸ்கி கூறிய குறைபாடுகள் நிவர்த்தி செய்யப்பட்டு மொழியில் ஆய்வுகள் கணிமொழியியல் ஆய்வாக மாறியது. உள்ளுணர்வு அடிப்படையிலான பல கருதுகோள்கள் தவறானவை என்பதை விரிதரவு ஆய்வு நமக்கு உணர்த்தியது.

விரிதரவுகள் மொழி வல்லுனர்களால் திறமையாகப் பயன்படுத்தப்படுகின்றது. அகராதிவியலார்கள், கலைச்சொல்லாக்க வல்லுனர்கள், கோட்பாட்டியலார்கள், பயன்பாட்டு மொழியலார்கள் தத்தம் செயல்பாடுகளுக்கு இன்றைய காலகட்டத்தில் விரிதரவைப் பெரிதும் பயன்படுத்துகின்றர். விரிதரவைக் கொண்டு அகராதிகள் உருவாக்கப்படுகின்றன. கலைச்சொல்லாக்கத்திற்காக விரிதரவு பெரிதும் உதவுகின்றது. கோட்பாட்டாளர்கள் தங்கள் கோட்பாடுகளை நிரூபிப்பதற்கும் தங்கள் கருதுகோள்களின் உண்மை நிலையை அறியவும் விரிதரவைப் பயன்படுத்துகின்றனர். பயன்பாட்டு மொழியிலார் பல மொழியில் பயன்பாடுகளுக்கு விரிதரவைப்

பயன்படுத்துகின்றனர். கணினி மொழியிலார்கள், கற்றல் கற்பித்தல் வல்லுனர்கள், சமூக மொழியலார்கள், உளமொழியியலார்கள், மானிட மொழியலார்கள், பண்பாட்டு மொழியியலார்கள் போன்றோர் விரிதரவை தங்கள் ஆய்விற்கும் பயன்பாட்டுக் கருவிகள் தயாரிக்கவும் பயன்படுத்துகின்றனர். பொருளடக்க வல்லுனர்களாக வரலாற்றியலார்கள், இலக்கிய திறனாய்வாளர்கள், சமூகவியலார்கள் விரிதரவைப் பயன்படுத்துகின்றனர். வரலாற்றியலார்களுக்கு விரிதரவு பல வழிகளில் கைகொடுக்கின்றது. விரிதரவைப் பயன்படுத்தி ஒரு நிகழ்வின் வரலாற்று காலம், இடம் போன்ற கூறுகளைச் சரியாகக் கணிக்கின்றனர். சமூகவியலார் வேறுபட்ட வகுப்பு, இனம், குடி சார்ந்த உண்மைகளையும் தகவல்களையும் விரிதரவின் உதவியால் சரியாகக் கணிக்கின்றனர். தகவல் வல்லுனர்களாக தகவல் மீட்பாளர்கள், கணினி மொழியியலார்கள், இயந்திர மொழிபெயர்ப்பாளர்கள், மொழியாய்வு ஆய்வாளர்கள் விரிதரவைப் பயன்படுத்துகின்றனர். தகவல் மீட்பாளர்கள் விரிதரவிலிருந்து தேவையான தகவல்களைப் பெற வேண்டி பல நுட்பங்களை உருவாக்கிச் செயல்படுகின்றனர். அவர்கள் அறிவு அடிப்படை உருவாக்கத்திற்கும் செயற்கை அறிவுக்கும் விரிதரவைப் பயன்படுத்துகின்றனர்.

கணினி மொழியியலார்கள் மொழியை ஆயவும் பகுப்பாயவும் முக்கியமான உபாயமாகச் செயல்படும் விரிதரவில் காணப்படும் புள்ளியியல் முறைமையுடன் தங்கள் செயல்பாட்டை ஒருங்கிணைக்க விரிதரவைப் பயன்படுத்துகின்றனர். மேலும் விரிதரவு தரவு மற்றும் அறிவடிப்படையின் மூலவளமாக, புள்ளியியல் உத்திகள் இலக்கண அடிப்படையில் ஆயப்பட்ட விரிதரவுகளின் விடுவரல்களின் மீது செயல்படுகையில் கூடுதல் திறனுள்ளதாய் மாறுவதன் காரணமாக மொழியில் உள்ள முறைமைகளின் இருப்பை அல்லது இல்லாமையைப் பரிசோதிக்கப் பயன்படுத்தப்படுகின்றது.

இயந்திர மொழிபெயர்ப்பாளர்கள் விரிதரவு ஒழுங்குமுறைகளின் உண்மையான திறனை அதிகரிக்க வேண்டி குறிப்பிடத்தக்கப் பங்களிப்பு செய்வதன் காரணமாகத் தேவையான மற்றும் பொருத்தமான மொழியியல் தகவல்களைப் பெறவும் அவர்களின் ஒழுங்குமுறைகளின் திறமையைப் பரிசோதிக்கவும் விரிதரவை நாடுகின்றனர்.

மொழி ஆய்விகளை உருவாக்கவேண்டி குறிப்புரை செய்யப்படாத மற்றும் குறிப்புரை செய்யப்பட்ட விரிதரவுகளைக் கூடுதலாகப் பயன்படுத்தப்படுவதால் மொழி ஆய்வாளர்கள் பல்வேறு வகைப்பட்ட விரிதரவுகளின் உருவாக்கத்தால் நன்மை அடைகின்றார்கள். விரிதரவு ஆய்வாளர்கள், தொழிநுட்பவியலார்கள், எழுத்தாளர்கள், அகராதியலார்கள், கல்வியாளர்கள், ஆசிரியர்கள், மாணவர்கள், மொழிகற்பவர்கள், அறிவாளிகள், வெளியீட்டாளர்கள் மற்றும் பிறர் உள்ளடங்கிய எல்லோருக்கும் பயனுள்ளதாக இருக்கின்றது.

துணை நூல்கள்

Aarts, J. and Meijs, W. (Eds.) 1984. *Corpus Linguistics: Recent Development in the Use of Computer Corpora in English Language Research*. Amsterdam-Atlanta, GA.: Rodopi.

Aarts, J. and Meijs, W. (Eds.) 1986. *Corpus Linguistics II: New Studies in the Analysis and Explanation of Computer Corpora.* Amsterdam-Atlanta, GA.: Rodopi.

Aijmer, K. and Allwood, J. (Eds.) 2004. *Dialogic Analysis: New Trends in Dialogue A*nalysis. Tubingen: Niemeyer.

Aijmer, K. and Altenberg, B. (Eds.) 1991. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman.

Altenberg B. and Granger, S. (Eds.) 2002. *Lexis in Contrast: Corpus-based Approaches*. Amsterdam/ Philadelphia: John Benjamins.

Aston, G. (Ed.) 2004. *Learning with Corpora*. Cambridge: Cambridge University Press.

Atwell, E. (Eds.) 1993. *Corpus Based Computational Linguistics*. Amsterdam: Rodopi.

Baker, M., Gill, F. and Tognini-Bonelli, E. (Eds.) 1993. *Text and Technology: In honour of John Sinclair.* Philadelphia: John Benjamins.

Barnbrook, G. 1998. *Language and Computers*. Edinburgh: Edinburgh University Press.

Béjoint, H. 2000. *Modern Lexicography: An Introduction*. Oxford: Oxford University Press.

Bernadini, S. 2000. *Competence, Capacity, Corpora*. Bologna: CLUEB.

Bernini G. (Ed.) 2001. *Pragmatic Organisation of Discourse in the languages of Europe*. Vol. 1. Berlin: Mouton de Gruyter.

Bharati, A., Chaitanya, V. and Sangal, R.1995. *Natural Language Processing: A Paninian Perspective*. New Delhi: Prentice-Hall.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus linguistics - Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Black, A.H. and Black, C.A. 1997. *A Conceptual Introduction to Morphological Parsing Using AMPLE System*. Dallas, Texas: Summer Institute of Linguistics.

Black, E., Garside, R. and Leech, G. (Eds.) 1993. *Statistically-driven Computer Grammars of English: the IBM/Lancaster Approach*. Amsterdam: Rodopi.

Bloom, L. 1970. *Language Development: Form and Function in Emerging Grammars*. Cambridge, Mass.: MIT Press.

Boguraev, B. and Pustejvsky, J. (Eds.) 1996. *Corpus Processing for Lexical Acquisition*. Cambridge, Mass.: MIT Press.

Botley, S.P., A.M. Mcenery, and A. Wilson (Eds.) 2000. *Multilingual Corpora in Teaching and Research*. Amsterdam-Atlanta, GA: Rodopi.

Bouillon, P. and Busa, F. (Eds.) 2001. *The Language of Word Meaning*. Cambridge: Cambridge University Press.

Brown, R. 1973. *A First Language: The Early Stages*. Cambridge, Mass.: Harvard University Press.

Burton, D.M. 1973. *Shakespeare's grammatical style: a computer-assisted analysis of Richard II and Anthony and Cleopatra*. Austin: University of Texas Press.

Butler, C.S. (Ed.) 1992. *Computers and Written Texts*. Oxford: Blackwell Publishers.

Carter, R. and Mccarthy, M. (Eds.) 1988. *Vocabulary and Language Teaching*. London: Longman.

Cassidy, F.G. Ed. 1985. *Dictionary of American Regional English*. Vol.1: Introduction and A-C. Cambridge, MA.: Harvard University Press.

Chomsky, A.N. 1968. *Language and Mind*. New York: Harcourt Brace.

Chomsky, A.N. 1972. *Studies in Generative Grammar*. The Hague: Mouton.

Coleman, J. and Kay, C.J. (Eds.) 2000. *Lexicology, Semantics and Lexicography: Selected Papers from the 4th G.L. Brook Symposium.* Amsterdam/ Philadelphia: John Benjamins.

Collins, P. and Blair, D. (Eds.) 1989. *Australian English*. St. Lucia: University of Queensland Press.

Crystal, D. 1995. *The Cambridge Encyclopaedia of the English Language*. Cambridge: Cambridge University Press.

Cuyckens, H. and ZAWADA, B. (Eds.) 2001. *Polysemy in Cognitive Linguistics.* Amsterdam/Philadelphia: John Benjamins.

Dash, N.S. 2005. *Corpus Linguistics and Language Technology (With Reference to Indian languages).* New Delhi: Mittal Publications.

Desai, S.K. 1974. *Experimentation with language in Indian Writing in English (Fiction).* Monograph of the Dept. of English, Shivaji University, Kohlapur, India.

Edwards, J.A. and Lampert, M.D. (Eds.) 1993. *Talking Data: Transcription and Coding in Discourse Research.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Eggins, S. 1994. *An Introduction to Systemic Functional Linguistics*. London: Pinter Publishers.

Flesch, R. 1946. *The Art of Plain Talk.* New York: Harper & Brothers.

Flowerdew, J. and Peacock, M. (Eds.) 2001. *Research Perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press.

Francis, W.N. and Kucera, H. 1964. *Manual of information to accompany A standard Corpus of present-day edited American English.* Dept. of Linguistics, Brown University, USA.

Francis, W.N. and Kucera, H. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston: Houghton Mifflin.

Fries, U., Mü, V. and Schneider, P. (Eds.) 1997. *From Aelfric to the New York Times*. Amsterdam: Rodopi.

Fuchs, C. and Vitorri, B. (Eds.) 1994. *Continuity in Linguistic Semantics*. Amsterdam and Philadelphia: John Benjamins.

Garside, R., Leech, G. and Mcenery, A. (Eds.) 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Garside, R., Leech, G. and Sampson, G. (Eds.) 1987. *The Computational Analysis of English: A Corpus Based Approach.* London: Longman.

Gerbig, A. 1997. *Lexical and Grammatical Variation in a Corpus: A Computer-Assisted Study of Discourse on the Environment.* London: Peter Lang Publishing.

Ghadessy, M., Henry, A. and Roseberry, R.L. Eds. 2001. *Small Corpus Studies and ELT: Theory and Practice.* Amsterdam/ Philadelphia: John Benjamins.

Gibson, H.N. 1962. *The Shakespeare Claimants: A Critical Survey of the Four Principle Theories Concerning the Authorship of the Shakespearean Play*. London: Methuen and Co.

Gove, P.B. (Ed.) 1961. *Webster's Third New International Dictionary of the English Language Sentences*. New York: Hartcourt, Brace and World.

Graf, D. 1996. *Relative Clauses in Their Discourse Context: A Corpus-Based Study*. Unpublished M.A. Thesis: Freiburg.

Granger, S. and Tyson, S.P. (Eds.) 2003. *Extending the Scope of Corpus-Based Research: New Applications, New Challenges*. Amsterdam: Rodopi.

Granger, S., Hung, J. and Tyson, S.P. Eds. 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Greenbaum, S. (Ed.) 1996. *Comparing English Worldwide: The International Corpus of English.* Oxford: Clarendon.

Greene, B. and Rubin, G. 1971. *Automatic Grammatical Tagging of English.* Technical Report. Department of Linguistics. Brown University, RI, USA.

Halliday, M.A.K. 1987. *Spoken and Written Modes of Meaning, Comprehending Oral and Written Language.* San Diego, CA: Academic Press.

Halliday, M.A.K. 1989. *Spoken and Written Language*. Oxford: Oxford University Press.

Halliday, M.A.K. and Hasan, R. 1976. *Cohesion in English*. London: Longman.

Halteren, H.V. (Ed.) 1999. *Syntactic Word Class Tagging*. Dordrecht: Kluwer Academic Press.

Hausmann, F.J., Reichmann, O., Wiegand, H.E. and Zgusta, L. (Eds.) 1990. *International Encyclopedia of Lexicography.* Vol.2. Berlin: Walter de Gruyter.

Herden, G. 1962. *Calculus of Linguistic Observation*. Hague: Mouton & Co.

Hickey, R. and Stanislaw, P. (Eds.) 1996. *Language History and Linguistic Modelling: A Festschrift for Jacek Fisiak.* Vol. 2. Berlin: Mouton de Gruyter.

Hoffman, C. 1955. *The Man Who was Shakespeare*. New York: Julias Messner Inc.

Hofland, K. and JOHANSSON, S. 1982. *Word Frequencies in British and American* English. Bergen: Norway Computing Centre for the Humanities.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hutchins, W. J. 1986. *Machine Translation: Past, Present, and Future.* Chichester: Ellis Harwood.

Ide, Nancy and Jean Véronis (Eds.) 1998. *Introduction to the special issue on word sense disambiguation: the state of the art*. Computational Linguistics. 24(1): 1-40.

Ilson, R.F. (Ed.) 1986. *Lexicography: An Emerging International Profession.* Manchester: Manchester University Press.

Ingram, D. 1989. *First Language Acquisition*. Cambridge: Cambridge University Press.

Jensen, J.T. 1990. *Morphology: Word Structure in Generative Grammar*. Amsterdam: John Benjamins.

Jespersen, O. 1909-1949. *Modern English Grammar on Historical Principles*. 7 Vols. London: Allen and Unwin.

Johansson, S. and Hofland, K. (Eds.) 1982. *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, S. and Stenström, A-B. (Eds.) 1991. *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyter.

Katamba, F. 1993. *Morphology*. London: Macmillan Press.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics.* New York: Addison-Wesley Longman Inc.

Kenny, A.J.P. 1982. *The Computation of Style*. Oxford: Pergamon Press.

Kettemann, C.B. and Marko, G. (Eds.) 2002. *Teaching and Learning by Doing Corpus Analysis. Language and Computers:* Studies in Practical Linguistics 42. Amsterdam-Atlanta, GA.: Rodopi.

Kilgariff, Adam and J. Palmer (Eds.) 2000. *Computer and the Humanities: Special Issue on Word Sense Disambiguation.* Vol. 34. No.1. 2000.

Kirk, J.M. (Ed.) 2000. *Corpora Galore: Analyses and Techniques in Describing English.* Amsterdam; Atlanta, GA: Rodopi.

Kruisinga, E. 1931-32. *A Handbook of Present-Day English*. Goringen: Noordhoff.

Kucera, H. and Francis, W.N. 1967. *Computational Analysis of Present Day American English*. Providence, RI: Brown University Press.

Kurath, H. Ed. 1954. *Handbook of the linguistic geography of New England*. Washington DC: American Council of Learned Societies.

Kytö, M., Ihalainen, O. and Rissanen, M. (Eds.) 1988. *Corpus Linguistics, hard and soft: Proceedings of the 8th International Conference on English Language Research on Computerised Corpora.* Amsterdam: Rodopi.

Lancashire, I., E. Carol, and C.F. Meyer (Eds.) 1997. *Synchronic Corpus Linguistics*. Bergen, Norway: ICAME.

Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Leech, G., G. Myers, and J. Thomas (Eds.) 1995. *Spoken English on Computer: Transcription, Markup and Applications*. Harlow: Longman.

Levy, M. 1997. *Computer Assisted Language Learning.* Oxford: Oxford University Press.

Ljung, M. (Ed.) 1997. Corpus-Based Studies in English. Papers from the 17th International Conference on English-Language Research Based on Computerized Corpora. Amsterdam: Rodopi.

Lock, A. (Ed.) 1978. *Action, Gesture and Symbol: The Emergence of Language*. London: Academic Press.

Lorge, I. 1949. *Semantic Content of the 570 Commonest English Words*. New York: Columbia University Press.

Macwhinney, B. 1991. *The CHILDES Project: Tools for Analyzing Tal*k. Hillsdale, N.J.: Lawrence Erlbaum.

Mair, C. and Hundt, M. (Eds.) 2000. *Corpus Linguistics and Linguistics Theory*. Amsterdam-Atlanta, GA: Rodopi.

Mcarthur, T. 1981. *Longman Lexicon of Contemporary English*. London: Longman.

Mccarthy, J. 1982. *Formal Problems in Semitic Phonology and Morphology*. New York: Garland.

Mccarthy, M. 1998. *Spoken Language and Applied Linguistics*. Cambridge: cambridge University press.

Mcenery, T. and Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mcenery, T., Rayson, P. and Wilson, A. (Eds.) 2002. *A Rainbow of Corpora: Corpus Linguistics and the Langauges of the World.* München: Lincom Europa.

Meyer, C.F. 2002. *English Corpus Linguistics*. Cambridge: Cambridge University Press.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. 1990. *Five Papers on WordNet. CSL Report 43,* Cognitive Science Laboratory, Princeton University, Princeton.

Miller, G.A. 1951. *Language and Communication*. New York: McGraw-Hills.

Mindt, D. 1995. *An Empirical Grammar of the English Verb: Modal Verbs*. Berlin: Cornelsen Verlag.

Mitkov, R. (Ed.) 2003. The Oxford Handbook of Computational Linguistiucs. Oxford: Oxford University Press.

Nelson, G., WALLIS, S. and AARTS, B. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English.* Amsterdam/ Philadelphia: John Benjamins.

Nihalani, P., TONGUE, R.K. and HOSALI, P. 1979. *Indian and British English: A handbook of Usage and pronunciation.* New Delhi: Oxford University Press.

Oakes, M.P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Ooi, V.B.Y. 1997. *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press

Oostdijk, N. and Dehann, P. (Eds.) 1994. *Corpus Based Research into Language*. Amsterdam-Atlanta, GA: Rodopi.

Orton, H. 1962. *Introduction to Survey of English Dialects*. Leads: E.J. Arnold and Son Ltd.

Palmer, H. 1933. *Second Interim Report on English Collocations.* Tokyo: Institute for Research in English Teaching.

Partington, A. 1998. *Patterns and Meanings - Using Corpora for English Language Research and Teaching.* Amsterdam/Philadelphia: John Benjamins.

Percy, C., Meyer, C.F. and Lancashire, I. (Eds.) 1996. *Synchronic Corpus Linguistics.* Amsterdam-Atlanta, GA: Rodopi.

Peters, B.P., Collins, P. and Smith, A. (Eds.) 2002. *New Frontiers of Corpus Research*. Language and Computers. Amsterdam-Atlanta, GA: Rodopi.

Pinker, S. 1995. *The Language Instinct: The New Science of Language and Mind*. Middlesex, England: Penguin Books Ltd.

Poutsma, H. 1926-29. *A Grammar of Late Modern English*. Groningen: Noordhoff.

PREYER, W. 1989. *The Mind of a Child*. New York: Appleton.

Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

Quirk, R., Gimson, A.C. and Warburg, J. 1968. *The Use of English*. 2nd Edition. London: Longman Group Ltd.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language.* London: Longman.

Ravin, Y. and Leacock, C. (Eds.) 2000. *Ploysemy: Theoretical and Computational Approaches*. New York: Oxford University Press Inc.

Schütze, H. 1997. *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*. Cambridge: Cambridge University Press.

Selkirk, E.O. 1983. *The Syntax of Wor*ds. Cambridge, Mass.: MIT Press.

Selting M. and Couper-KUHLEN, E. (Eds.) 2001. *Studies in Interactional Linguistics*. Amsterdam/ Philadelphia: John Benjamins.

Sharman, R. 1990. *Hidden Markov model methods for word tagging*. Report 214. Winchester: IBM UK Scientific Centre.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Smith, L. (Ed.) 1981. *English for Cross-cultural Communication*. London: Macmillan.

Souter, C. and Atwell, E. (Eds.) 1993. *Corpus Based Computational Linguistics.* Amsterdam: Rodopi.

Spencer, A. 1991. *Morphological Theory*. Oxford: Basil Blackwell.

Sperberg-Mcqueen, C.M. and Burnard, L. 1994. *Guidelines for Electronic Text Encoding and Interchange.* Chicago and Oxford: ACH-ACL-ALLC Text Encoding Initiative.

Stenström, A-B, Andersen, G. and Hasund, I.K. 2002. *Trends in Teenage Talk: Corpus Compilation, Analysis and Findings.* Amsterdam: John Benjamins.

Stern, W. 1924. *Psychology of Early Childhood up to Six Years of Age*. New York; Holt.

Stubbs, M. 1996. *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell. Publishers.

Summers, D. 1991. *Longman/Lancaster English Language Corpus: Criteria and Design*. Harlow: Longman.

Svartvik, J. (Ed.) 1990. *The London Corpus of Spoken English: Description and Research*. Lund Studies in English 82. Lund: Lund University Press.

Svartvik, J. (Ed.) 1992. *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 - Stockholm,* 4-8 August 1991. Berlin, New York: Mouton De Gruyter.

Tannen, D. (Ed.) 1982. *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, New Jersey: Ablex Publishing Corporation.

Thomas, J. and SHORT, M. (Eds.) 1996. *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*. London and New York: Addison Wesley Longman.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Uta, F. (Ed.) 1979. Cognitive Processes in Spelling. London: Academic Press.

Vera, D.E.J. (Ed.) 2002. *A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics.* Amsterdam: Rodopi.

Véronis, J. (Ed.) 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora*. Dordrecht: Kluwer Academic Publishers.

Weigand, E. and Dascal, M. (Eds.) 2001. *Negotiation and Power in Dialogic Interaction*. Amsterdam/Philadelphia: John Benjamins.

Wichmann, A., Fligelstone, S., Mcenery, T. and Knowles, G. (Eds.) 1997. *Teaching and Language Corpora.* London: Longman.

Wills, J.D. 1990. *The Lexical Syllabus*. London: Collins.

Winograd, T. 1983. *Language as a Cognitive Process*. Vol. I. Mass.: Addison-Wesley.

Wiseby, R.A. (Ed.) 1971. *Computer in Literary and Linguistic Research*. *Papers from the Cambridge Symposium*. Cambridge: Cambridge University Press.

Young, S. and G. Bloothooft (Eds.) 1997. *Corpus-Based Methods in Language and Speech Processing.* Vol-II. Dordrecht: Kluwer Academic Press.

YULE, G.U. 1964. *The Statistical Study of Literary Vocabulary.* Cambridge: Cambridge University Press.

Zernik, U. (Ed.) 1991. *Lexical Acquisition*. Englewood Cliff, NJ: Erlbaum.

Zipf, G.K. 1936. *The Psycho-biology of Language: An Introduction to Dynamic Philology*. London: G. Routledge.

Zipf. G.K. 1949. *Human Behaviour and the Principle of Least Effort: An Introduction of Human Ecology*. Cambridge, Mass.: Addison-Wesley.